# NATIONAL CENTER FOR EDUCATION STATISTICS

## Working Paper Series

The Working Paper Series was created in order to preserve the information contained in these documents and to promote the sharing of valuable work experience and knowledge. However, these documents were prepared under different formats and did not undergo vigorous NCES publication review and editing prior to their inclusion in the series.

**U. S. Department of Education**
**Office of Educational Research and Improvement**

# NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

# Methodological Issues in the Study of Teachers' Careers: Critical Features of a Truly Longitudinal Study

Working Paper No. 96-01 January 1996

Contact: Dan Kasprzyk
Education Surveys Division
(202) 219-1588

**U. S. Department of Education**
**Office of Educational Research and Improvement**

**U.S. Department of Education**
Richard W. Riley
Secretary

**Office of Educational Research and Improvement**
Sharon P. Robinson
Assistant Secretary

**National Center for Education Statistics**
Jeanne E. Griffith
Acting Commissioner

**Elementary/Secondary Education Statistics Division**
Paul D. Planchon
Associate Commissioner

**National Center for Education Statistics**

The purpose of the Center is to collect and report "statistics and information showing the condition and progress of education in the United States and other nations in order to promote and accelerate the improvement of American education."—Section 402(b) of the National Education Statistics Act of 1994 (20 U.S.C. 9001).

**January 1996**

## Foreword

Each year a large number of written documents are generated by NCES staff and individuals commissioned by NCES which provide preliminary analyses of survey results and address technical, methodological, and evaluation issues. Even though they are not formally published, these documents reflect a tremendous amount of unique expertise, knowledge, and experience.

The ***Working Paper Series*** was created in order to preserve the information contained in these documents and to promote the sharing of valuable work experience and knowledge. However, these documents were prepared under different formats and did not undergo vigorous NCES publication review and editing prior to their inclusion in the series. Consequently, we encourage users of the series to consult the individual authors for citations.

To receive information about submitting manuscripts or obtaining copies of the series, please contact Suellen Mauchamer at (202) 219-1828 or U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 555 New Jersey Ave., N.W., Room 400, Washington, D.C. 20208-5652.

Susan Ahmed
Acting Associate Commissioner
Statistical Standards and
Methodology Division

Samuel S. Peng
Statistical Service and
Methodological Research

*This page intentionally left blank.*

# Methodological Issues in the Study of Teachers' Careers:

# Critical Features of a Truly Longitudinal Study

Judith D. Singer
John B. Willett

January 1996

Harvard University Graduate School of Education
Appian Way
Cambridge, MA 02138

*This page intentionally left blank.*

## Table of Contents


Page

Foreword . . . . . iii

Table of Contents . . . . . vii

1.0 Introduction . . . . . 1

2.0 Research context and goals: Why should NCES conduct a new longitudinal study of teachers' careers? . . . . . 4
2.1 What is currently known about teachers' careers . . . . . 4
2.2 Gaps in our knowledge of teachers' careers . . . . . 5
2.2.1 Teacher quality . . . . . 6
2.2.2 Teachers' work contexts . . . . . 6
2.2.3 Teachers' worklives . . . . . 7
2.2.4 Teachers' career paths . . . . . 7

3.0 Implications of the Research Goals: Six General Principles of Research Design . . . . . 9
3.1 Principle #1: NCES must collect truly longitudinal data . . . . . 9
3.2 Principle #2: NCES must view "time" as both an outcome *and* a predictor . . . . . 12
3.3 Principle #3: NCES must collect data on both time-varying and time-invariant measures . . . . . 15
3.4 Principle #4: NCES must collect data prospectively when necessary . . . . . 16
3.5 Principle #5: NCES must collect data beginning in multiple base years . . . . . 17
3.6 Principle #6: NCES must collect data from at all relevant levels of the organizational hierarchy . . . . . 20

4.0 Identifying the target population: Whom should NCES study? . . . . . 22
4.1 Diagramming the teaching career . . . . . 22
4.2 What *do* we know about the stock of current teachers? . . . . . 28
4.2.1 Number of spells . . . . . 28
4.2.2 Years of service during each spell . . . . . 30
4.3 Four possible target populations . . . . . 31
4.3.1 Stock sample of teachers . . . . . 32
4.3.2 First year teacher study . . . . . 36
4.3.3 Two other options: A beginning teacher study and a stratified stock sample . . . . . 38
4.4 What is our recommendation? . . . . . 40

5.0 The time dimensions of data collection: How often, and for how long, should data be collected? 44
5.1 How many waves of data are required? . . . . . 45
5.1.1 The shape of the individual growth trajectory . . . . . 45
5.1.2 The precision of estimates of growth . . . . . 48
5.1.3 The reliability with which change can be measured . . . . . 49

5.2 For how long should teachers be observed? . . . . . 53
5.3 What is our recommendation? . . . . . 60

6.0 Issues of implementation: Implications of other methodological concerns for the conduct of the longitudinal study . . . . . 61
6.1 Replicating the study in multiple base years . . . . . 61
6.2 Modes of data collection . . . . . 62
6.3 Retrospective reconstruction of event histories . . . . . 64
6.4 Ensuring measurement equatability . . . . . 67
6.5 Tracking teachers in and out of schools . . . . . 70
6.6 Embedding substudies of natural experiments in the larger study . . . . . 72
6.7 Pilot studies . . . . . 74

7.0 Conclusion . . . . . 75

8.0 References . . . . . 76

## 1.0 Introduction

The National Center for Education Statistics (NCES) is exploring the possibility of conducting a large-scale multi-year study of teachers' careers. Unlike current NCES surveys of teachers--the cross-sectional Schools and Staffing Survey (SASS) and its companion one-year prospective Teacher Followup Survey (TFS)--the proposed new study is intended to follow a national probability sample of teachers over an extended period of time. Recognizing that this effort will involve considerable expense and time, NCES is in the process of evaluating the need for, and the feasibility of, such a data collection enterprise.

As part of this process, NCES commissioned a panel of experts to present their perspectives on the *substantive* issues involved in a longitudinal study of teachers' careers. Taken together, these four papers (Billingsley, 1992; Grissmer & Kirby, 1992; Theobold & Gritz, 1993; and Weiss, 1992) substantiate the need for a longitudinal study and argue for its implementation during the late 1990s.

At a January 1993 planning conference in Washington DC attended by these experts and other researchers and policymakers from both within and outside of NCES, however, it became clear that there were a number of *methodological* issues that needed to be addressed before initiating the study. Precisely which teachers should be studied? Should NCES follow a probability sample of all teachers at varying points in their careers, or should they focus on teachers at a particular career juncture, say beginning teachers, mid-career teachers, or perhaps teachers nearing retirement? For how long should the teachers be followed? Is complete follow-up until retirement necessary or would a shorter time-period suffice? How often should teachers be contacted? Is it necessary to collect data every year, or is a shorter interval necessary or a longer interval sufficient?

After the planning conference, NCES asked the current authors to make recommendations focusing on these and other methodological issues that arise in the design and implementation of a longitudinal study of teachers' careers. The current paper, the result of that request, builds on the conclusions of the January 1993 planning conference, our subsequent discussions with NCES staff, and our previous work on both research design and teachers' careers.

Of course, most issues of longitudinal research are inexorably linked, by cost if nothing else. If individuals are to be measured more frequently, for example, it may be necessary to include fewer respondents in the sample in order to limit the overall cost of the project. What constitutes adequate measurement for one research purpose may be inadequate for another. As a consequence, there may be no single optimal design that is best suited to research on all aspects of a teacher's career. Design decisions have to be linked to arguments about competing research goals, and invariably, compromises must be struck.

Rather than trying to offer a single "optimal" design in this report, we have decided to offer, and critique, a variety of alternative designs. We begin (in Section 2), by reviewing the research context and goals of the proposed study, identifying four principal substantive gaps in the current knowledge base on the teaching career. In the following sections, we offer both *general* and *specific* recommendations on the design of the proposed study. In Section 3, we make general proposals, identifying what we consider to be a set of important design principles that flow from the research goals. In Sections 4, 5, and 6, we follow through on these general principles by providing concrete recommendations for specific aspects of the proposed design. We focus particularly on designating the target population and the specific sample to be tracked (Section 4), as well as on the length and periodicity of data collection (Section 5). In addition, we comment on implementation issues that will arise as NCES expands its data collection efforts beyond the current cross-sectional and one-year follow-up studies of teachers (Section 6). By beginning with general principles instead of specific recommendations, we hope that if NCES decides that, for reasons of cost or scope, it cannot adhere to our specific recommendations, the general

principles may still influence the shape of any alternative research design.

In identifying the major methodological issues involved in the design of a longitudinal study of teachers' careers, we have purposefully omitted discussion of several common design issues, except where they impact our focal topics. In particular, we do not discuss statistical power and sample size, nor do we discuss the practical issues involved in listing the target population and drawing the desired probability sample. We do not consider these topics unimportant but we believe that NCES can best address these practicalities *after* dealing with the larger conceptual issues outlined here.

## 2.0 Research context and goals:

## Why should NCES conduct a new longitudinal study of teachers' careers?

During the past decade, as quantitative research on teachers' careers has accumulated, its quality has also improved. The empirical research literature, once comprised solely of aggregate annual attrition rates computed at district and state levels (see, e.g., Darling-Hammond, 1984), now includes large-scale in-depth studies of the occurrence and spacing of critical events in the teaching career: who stays, who leaves, who returns, and when and why these transitions occur (Willett & Singer, 1991). As we contemplated the design and implementation of a new longitudinal study, we asked ourselves what the research community has learned from this expanding literature, what were the knowledge gaps that remain unfilled, and what contribution a new NCES data collection effort might make to the field.

### 2.1 What is currently known about teachers' careers?

Although there have been many intensive small scale studies of teachers in individual school districts, the most generalizable information on teachers' careers has made use of data from one of four sources:

- *state-level administrative records*, which are used to reconstruct the longitudinal career histories of entire cohorts of newly-hired teachers in a single geographic area (see, e.g., Grissmer & Kirby, 1992; Murnane, Singer, & Willett, 1988, 1989; Schlechty & Vance, 1981; Theobold, 1990);
- *longitudinal surveys of national probability samples of college graduates*, which are used to construct the longitudinal career histories of individuals who ever taught anywhere in the United States (see, e.g., Hafner & Owings, 1991; Heyns, 1988; Murnane, Singer, Willett, Kemple, & Olsen, 1991);

- *retrospective reports from national probability samples of current teachers* who participate in NCES' tri-annual SASS, which are used to describe current teachers' previous career decisions and their future intentions (see, e.g., Bobbitt, Faupel, & Burns, 1991);
- *one-year followup reports from national probability subsamples of teachers* who participate in the NCES' tri-annual TFS, which are used to estimate annual attrition rates, to identify the career decisions of teachers who participated in SASS, and to describe teachers' current work conditions (see, e.g., Choy, Medrich, Henke, & Bobbitt, 1992).

What have we learned from these databases? Although a comprehensive review is beyond the scope of the present paper, several key findings have emerged (for recent reviews, see the National Academy of Sciences 1992 report *Teacher Supply, Demand, and Quality: Policy Issues, Models and Data Bases* and Darling-Hammond & Sclan, in press). We know, for instance, that the appeal of teaching declined precipitously during the 1970s and 1980s. We know that the first years in teaching continue to be the riskiest for all teachers. Schools still lose those teachers who score particularly well on standardized tests, those who have attractive career opportunities outside the schools, or those who are paid comparably lower salaries. A further lesson has been the great mobility of the US teaching force: not only do many teachers who leave teaching ultimately return, but it is the reserve pool of former teachers--not the pool of recent college graduates--who now comprise the major source of teacher supply.

## 2.2 Gaps in our knowledge of teachers' careers

Despite the fact that research has moved beyond the simple attrition rate and that a deeper understanding of the relationship between teacher supply and demand has developed, there is still much to be learned. Few large-scale studies have asked *why* teachers behave the way they do. In particular, we believe that there are four prominent areas worthy of further investigation: (a) teacher quality, (b) teachers' work contexts, (c) teachers work lives, and (d) teachers' career paths.

**2.2.1 Teacher quality**. No large scale study of teachers' careers has yet attempted to measure the "quality" of the nation's teaching force. It is only by measuring this elusive construct that we can begin to ask whether the teachers who leave the schools are perhaps the very ones we may wish to leave--those of lower quality. In this paper, we will not wade into the substantive, political, and methodological controversies surrounding the measurement of teacher quality (Kennedy, 1992). We address the topic only through its ramifications for other aspects of design, as when we discuss the need for data collected from individuals other than the teachers themselves. Because of the importance of the topic, however, we recommend that NCES contact experts in the field in order to explore possible strategies for collecting valid and reliable data on teacher quality. The state of knowledge about teachers' careers is such that the issue of "quality" is increasingly emerging as a fundamental question (Shulman, 1992). To conduct a longitudinal study of teachers' careers without measuring some aspects of teacher quality would be, to our mind, a grave omission.

**2.2.2 Teachers' work contexts**. The current SASS and TFS gather only limited data on the contexts of teachers' work. Administrators, principals, and colleagues respond to questionnaires, but they are asked only general questions about their schools, not specific questions about the conditions under which specific surveyed teachers work. Teachers, too, are asked to describe their perceptions of administrators and colleagues but, with the exception of the administrator questionnaire, direct linkage is difficult, if not impossible. A new longitudinal study would provide an ideal setting in which to delve further into this topic: Who are the administrators with whom teachers work? How do they view the teachers? Such linkage would expand our knowledge about *who* is leaving teaching, and who returns, and *why*. A second key element of the teacher work context is the students that the teachers serve. A longitudinal study of teachers' careers would provide an ideal setting in which to gather data on students in the classes of surveyed teachers. It is true that data on teachers has been gathered in other student-centered NCES data collection efforts (e.g., NLS-72, HSB, NELS) but the students have been retained in

the sample across waves while the teachers have come and gone. In a longitudinal study of teachers, the flip side would be the case--teachers would be retained in the sample across waves while the students came and went. This would allow researchers to determine, for example, whether fluctuations in teachers' commitment to the profession are related to variation in the students they serve. Perhaps teachers leave teaching when they have to teach more difficult students.

**2.2.3 Teachers' worklives.** A third fundamental knowledge gap in the study of the US teaching force concerns an understanding of teachers as workers, leaders, mentors, and individuals and how these roles evolve and change over time. The currently available longitudinal data sources (primarily administrative records and large multi-purpose national surveys) have not included the kinds of information that researchers need to fully describe teachers' lives in schools. And although NCES' two current data collection efforts on teachers provide some insight into these topics, these studies are not truly longitudinal, thereby describing status, and not change. The clear consensus from the January planning conference was that a major contribution of a new longitudinal study would be the collection of detailed data on teachers lives in school. In particular, the study should gather data about teachers' work roles and working conditions and how these features vary across settings and change over time (Billingsley, 1992). Under this rubric falls the study of topics such as the on-going support and development of teachers as professionals and leaders; teachers' perceptions of their work climate, reform initiatives, school administrators, and operations; and teachers' job commitment and job satisfaction. Although the current SASS and TFS attempt to gather measures of some of these constructs, the questions are brief, primarily because the current studies focus on issues such as degree attainment and certification. By gathering longitudinal data on teachers' worklives, researchers could determine *which* teachers are most likely to leave, and whether certain characteristics of jobs and schools are associated with lower attrition, higher satisfaction, and stronger commitment.

**2.2.4 Teachers' career paths.** Fourth, despite the fact that there has been more research on this aspect of teachers' careers than on any other, substantial knowledge gaps persist. Although most

researchers recognize that teachers' careers unfold over time, most studies still rely on data collected retrospectively, cross-sectionally, or at only two points in time. No study has yet to juxtapose career decisions alongside full data on wage and benefit histories as well as workplace conditions and family demands. No study has attempted to track the labor market experiences of current teachers, former teachers, and returning teachers. No study has attempted to follow newly licensed teachers as they enter teaching, leave teaching, return to teaching, and perhaps leave once again. As we argue elsewhere (Willett & Singer, 1991), it is only through the tracking of teachers along their various extended career paths that we will understand when and why teachers make the career transitions that they do.

The need for data to fill these four gaps in teacher quality, the context of teaching, teachers' worklives, and teachers' career paths are at the core of our substantive recommendations to NCES. Of course, these are not the only four areas that might be addressed. We have highlighted them because we believe they represent the four most important foci for a future longitudinal study and because, taken together, they have direct implications for research design. When we note, for example, that NCES should collect data on the context of teaching, this implies that data be collected from the people with whom teachers work. Because of the direct link between research goals and research design, we now discuss the design implications of these four substantive goals.

## 3.0 Implications of the Research Goals: Six General Principles of Research Design

To design a study that will provide data suitable for addressing the four research domains outlined in the previous section, we believe that NCES must adhere to a set of six general design principles, which we outline in this section. In outlining these general principles, we establish the core ideas that will underpin our specific proposals for the design of the future longitudinal study of teachers' careers.

### 3.1 Principle #1: NCES must collect truly longitudinal data

Few studies of teachers' careers have actually followed the careers of teachers over extended periods of time. Instead, much of what is known about teachers' careers has been constructed piecemeal by linking together the findings of studies with cross-sectional, or limited longitudinal, designs. Our fundamental position is that it is *changes* in teachers and *transitions* in their careers that should be the critical substantive focus of the proposed new research and that, unless decent longitudinal data are collected, researchers will be unable to study these important changes carefully and effectively.

Why are longitudinal data required for the measurement of change? Because, if an analysis of cross-sectional data, from the SASS say, reveals that teachers with more years of experience have lower levels of commitment to the profession than teachers with fewer years of experience, we cannot necessarily infer that teachers' commitment decreases over time as they accumulate experience. In a cross-sectional study, the teachers with more years of experience differ in important and fundamental ways from the teachers with fewer years of experience--they graduated from college in different years, they were licensed in different years, they entered teaching in different years, and they have taught under different types of working conditions. Differences in commitment may be due to these differences in

background characteristics, attributes and conditions, rather than to any systematic decrease in individual commitment over time. Information about status in any given year does not reveal anything about the process that lead the teachers to this particular status at this particular time. Perhaps the commitment of many mid-career teachers has actually *increased* over time, and the low level we see in a cross-sectional survey is *higher* than we would have seen had we collected data on these teachers during their previous years on the job. If we do not collect longitudinal data, we will never know!

Recognizing these limitations of cross-sectional data, the current NCES data collection program includes the Teacher Follow-Up Survey (TFS), conducted one year after each base year SASS. Although the TFS is helpful for estimating an annual attrition rate and for studying the short-term mobility of the teaching force, two waves of data are inadequate for studying change in teacher attributes over time.[1] Two waves of data separated by only one school year cannot be used to effectively portray the complex patterns of growth and change that we expect teachers to display over their careers. Among the many problems with two-wave designs is their failure to collect sufficient data to characterize the *shape* of each teacher's growth trajectory. Two waves of data tell researchers only about each teacher's status at two points in time; there is no information about *how* the teachers got from point A to point B. In that single year, did all the change occur immediately after the beginning of the school year? Did most teachers change at a steady pace, at equal amounts each month? Did some teachers remain at a steady pace for most of the year, only to fall off in commitment by the spring? What will happen as these teachers continue to teach? With only two waves of data separated by only one school year, it is impossible to know.

To measure change over time, data must be collected at more than two timepoints. As more waves of data are collected, the researcher becomes increasingly able to construct a more finely-grained picture of the development of teacher attributes over time. Patterns of change may even be seen in

---

① We use the word change here in its broadest sense, to encompass changes in such diverse domains as educational status, family status, employment status, and attitudes, knowledge, and behavior.

simple plots of status versus time. Is change linear or curvilinear? Do most teachers peak after their first few years on the job, or do many increase steadily over time? Are increases in one dimension (say, self-perceived competency) accompanied by decreases in other dimensions (say, job stress)? Are teachers with declining levels of efficacy more likely to leave than those on an upward trajectory? Do most changes occur during the early years on the job and the years near retirement, or are there substantial change during mid-career as well?

In designing a new longitudinal study of the teaching career, we must be concerned not simply with the number of waves of data collected but also with the duration of data collection. The duration of the proposed longitudinal study must be long enough to permit careful study of the changes and transitions that are occurring in teachers' lives over time. Teachers are not students; the changes they exhibit are likely to be subtle and slow. Attrition rates are low. Yes, teachers move between schools and school districts, there are pockets of great turnover and change, and there are increasing percentages of teachers nearing retirement, but as a whole, in comparison to years past, the nation's teaching force has become relatively stable.

The study of change in a relatively stable environment requires the collection of longitudinal data over a longer period. In addition, it is only when such long-duration empirical records become available that researchers can address important policy questions about the return of former teachers to teaching. The reason is simple: if a longitudinal study is to describe effectively the return of former teachers to the profession, then data must be collected over a sufficiently long time-period to permit an analyzable subsample of teachers to return.

If there were no need for immediate results and cost were no object, we might recommend that NCES continue to observe teachers data throughout the *entirety* of their teaching careers--from choice of major in college through eventual retirement from the profession. Such a "womb to tomb" data collection is ideal for studying change over the whole career. It would allow us to learn, for example, whether career patterns and levels of commitment differ for teachers who appeared committed to an

education career in their college years in comparison to other teachers who came to the profession at a later point in time. Or whether teachers we might have been able to identify, on the basis of early signs and signals, were likely to leave teaching, never to return.

Given a sufficiently frequent set of temporal observations, womb-to-tomb data collection has another important appeal--it enables the researcher to observe *change as it occurs.* Current information on teachers' careers provides the research community with only the most limited information about when and why changes occur. If NCES were able to collect data frequently throughout the entire teaching career, then crucial times of change or transition would be less likely to be missed. And, the record of time-varying longitudinal information available on the teachers' careers prior to the transition in question could then be used to predict transition occurrence and spacing (Willett & Singer, 1993).

While not in the position to collect unlimited amounts of longitudinal data over the entire teacher career, NCES must recognize that answers to important questions about changes and transitions in teachers' worklives, quality, work contexts, and career paths require more waves of data than the agency currently collects. Consequently, we recommend that NCES expand their longitudinal data collection in two ways: (1) by collecting more waves of data; and, (2) by extending the duration of data collection over a longer period of time. In Section 5, we use data from the SASS and the TFS to make these recommendations more specific.

### 3.2 Principle #2: NCES must view "time" as both an outcome *and* a predictor

Most longitudinal studies of teachers' careers have viewed chronological time as either an outcome or as a predictor.[2] This bifurcation has arisen as an artifact of both disciplinary boundaries and substantive focus. Studies that have examined transitions in and out of teaching have arisen primarily out of an economic tradition, and have viewed time as the conceptual outcome, an object of study in its

---

② See, for example, the special issue of the *International Journal of Educational Research* focused on teachers' careers edited by Huberman, 1988.

own right. Such researchers examine *whether* the teachers experience particular transitions (entering teaching, moving to another school or school district, leaving teaching, or returning to teaching) and *when* these transitions occur. On the other hand, studies that examined changes in teachers' attitudes, knowledge, or behavior over time have had their origins in disciplines such as sociology and psychology and, in contrast, have treated time as a predictor. Researchers in this tradition study *whether* and *how* the attributes of teachers change over time.

We believe that both perspectives deserve equal voice if an effective longitudinal study of teachers' careers is to be designed. It would be an error, we believe, for NCES to prioritize these perspectives, emphasizing the investigation of the timing of transitions in and out of teaching (time as an outcome), say, over the study of changes in teacher attributes over time (time as a predictor). In the past decade, there has been a great deal of policy interest in issues of teacher supply and demand. Hence, studies that have treated time as the conceptual outcome have received more attention than have studies that have treated time as a predictor. But at the January 1993 planning conference, Emerson Elliot, Commissioner of NCES stressed that increasing attention needed to be given to the study of changes in teachers worklives over time.

Decisions about number of waves, spacing of waves, and length of data collection are usually made within the context of only one of these perspectives. Researchers interested in measuring change, for example, treat time as a predictor and make design recommendations based on this view (see, e.g., Willett, 1988). The goal of design, from this perspective, is to determine the length of data collection and the spacing of waves so that a precise and unbiased statistical summary of *change* can be obtained as efficiently as possible. Researchers interested in measuring event occurrence, in contrast, treat time as an outcome and make design recommendations from that point of view (see Singer & Willett, 1991). The goal of design, from this perspective, is to determine the length of data collection and the spacing of waves so that a statistical summary of *event occurrence* can be computed as precisely as possible. As described in Section 5, on occasion, both perspectives lead to the same design recommendation; on other

occasions, however, they conflict.

It is our view that the design of a multi-purpose longitudinal study of teachers' careers must not favor one perspective over the other. The practical implication of this view is that NCES may have to commit to collecting data more frequently and for a longer period of time than they might have chosen had they decided on designing this longitudinal study from only one perspective. If, for example, the "time as an outcome" perspective would allow a bi-annual data collection schedule, but the "time as a predictor" perspective points to a *semi-annual* schedule, we would recommend the more frequent periodicity. Our reason is simple. Use of the more frequent periodicity would certainly not harm researchers working in the first tradition, but failing to do so would certainly stymie researchers operating in the second tradition.

Further reflection on this topic also suggests that researchers from the two traditions may not be as nearly at odds with each other as one might initially suspect. We began our discussions of design by considering these two perspectives as separate and distinct. Even a simple review of the literature reinforces this stereotype. Yet we now see the possibility of a merger between traditions because of a simple insight: one tradition's outcome is the other tradition's predictor. Those attributes in which one researcher is interested in measuring change over time (teacher self-esteem, relationship with principals) are exactly those attributes that another researcher might consider as time-varying covariates in the study of the timing of events in a teacher's life. The problem historically has been that studies in this tradition have not collected data from the teachers themselves, precluding the investigation of such effects. Conversely, the events whose occurrence researchers are studying (transferring from one school to another) may affect a teacher's growth trajectory, producing important and measurable impacts on the level, shape or curvature. But here, too, the unavailability of longitudinal records on event occurrence has precluded the investigation of these sorts of effects.

Our review of the literature suggests that lack of relevant data has allowed these two research traditions to grow in isolation. A major contribution of a truly longitudinal study of teachers' careers

would be the potential synnergy that would come from researchers working in both these traditions using the same data resource. We recommend that NCES grasp this opportunity with both hands.

### 3.3 Principle #3: NCES must collect data on both time-varying and time-invariant measures

All of the data collected in a longitudinal study can be classified as either time-invariant or time-varying. As their label suggests, time-invariant variables measure characteristics of individuals that do not change over time. In the study of teachers careers, important time-invariant variables include demographic characteristics such as year of birth, race/ethnicity, sex, year of college graduation, year of licensure, and college major. Data on time-invariant measures need be collected only *once* during a longitudinal study, such as in the base year of the survey. This frees subsequent data collection time allowing it to be devoted to the acquisition of time-varying information.

Time-varying variables are also true to their name--they have values that vary over time. In the study of teacher careers, most of the variables of greatest research interest are time-varying: teacher efficacy, class size, salary, working conditions, family composition, and so forth. Data on time-varying predictors must be collected systematically and repeatedly over time. Although it is possible to reconstruct the values of some time-varying measures retrospectively, data will be gathered with greater validity and precision if the values of time-varying measures are gathered *as they fluctuate*. Can a researcher reasonably expect a teacher with ten years of teaching experience to reliably and validly retrospect back to his or her experiences, opinions and feelings during the first year in the classroom? If researchers need to investigate the chronological variation of time-varying measures (as in studies of change), or if they want to use time-varying measures as predictors in studies of whether and when teachers transfer from a school or leave teaching, NCES must measure these variables as their values fluctuate over time.

This measurement issue has a direct implication for the spacing of the waves of data collection. If the values of time-varying measures fluctuate rapidly, then waves of data collection must be spaced

closely. If they are spaced too far apart, important fluctuations will be missed or teachers will have to retrospect to "fill in the blanks." On the other hand, if the values of time-varying variables change less rapidly, then waves of data collection can be spaced further apart. After all, if responses change infrequently, the measure becomes essentially "time-invariant" during the inter-interview periods.

In the design of a large-multi-purpose study where many constructs are being measured simultaneously, the values of some important time-varying predictors will fluctuate frequently (teacher efficacy, for example, may change on a weekly or even a moment-by-moment basis), while the values of others will fluctuate relatively infrequently (class size or type of students served, for example, may change only on a semester or annual basis). Thus, NCES might consider using different data collection periodicities for different measures, with selection of periodicity being determined via pilot study or by expert opinion. It may be cost-effective and efficient, for example, to collect some types of time-varying data on a semi-annual basis, and other types of data on an annual or bi-annual basis. In a later section of this paper, we discuss the implications of these issues for the spacing of waves of data-collection. For now, however, we simply note that NCES must commit adequate resources to measuring the values of time-varying variables as often as is necessary.

### 3.4 Principle #4: NCES must collect data prospectively when necessary

From one perspective, nearly all of the data that NCES collects on its teacher questionnaires are retrospective. When a teacher is asked to rate whether student violence is a problem in his or her school, the teacher is reflecting (retrospecting) on his or her experience. A researcher can attempt to limit the time frame involved in the retrospection by adding temporal boundaries to the question. In some instances, an item might be worded "*During the past week,...*" whereas in others, the item might begin *"During this school year...."*

As is well known, retrospective data collection is fraught with problems. First consider the easiest type of data to collect via retrospection--data on the occurrence and spacing of critical events.

Although rare and important events--such as college graduation, first teaching job, entering this school district--may be remembered indefinitely, and highly salient events--such as a leave of absence from teaching--may be remembered for several years, habitual events--such as daily work activities--are forgotten almost immediately (Bradburn, 1983; Sudman & Bradburn, 1982). The more extreme the length of retrospection, the greater are the errors committed. Three types of errors are especially common: (a) *memory failures*, in which respondents forget events entirely; (b) *telescoping*, in which events are remembered as having occurred more recently than they actually did; and (c) *rounding*, in which respondents drop fractions and report even numbers or numbers ending in 0 and 5. These errors create different biases: Memory failures lead to underreporting, telescoping to overreporting, and rounding to both.

If gathering retrospective information about event occurrence is difficult, gathering retrospective data that require qualitative or quantitative judgments is next to impossible. When you ask teachers to retrospect about "states"--their attitudes towards their jobs, efficacy, satisfaction, or commitment--in years gone by, then the errors can only escalate. It is virtually impossible to collect retrospectively reliable and valid restrospective information on attitudes and affect, especially if these attributes are fluctuating with time.

The implication of this recognition is clear: NCES must commit itself to collecting prospective data whenever necessary in order to ensure the validity and reliability of responses. As we discuss in Section 5, adherence to this principle may require increased periodicity of data collection for certain types of information. Data may be collected retrospectively only when their reliability and validity are not challenged.

### 3.5 Principle #5: NCES must collect data beginning in multiple base years

In its original plan for the SASS, NCES wisely recognized that the teaching force changes over time. The decision to field a national survey every three to four years (with a smaller follow-up after

each round of data collection) ensures that time-related differences in the teaching force can be registered. By comparing the proportion of fully certified teachers in 1987 with the proportion of fully certified teachers in 1990 and 1994, researchers can comment cogently on historical changes in the teaching force from year to year. This issue is of great interest to policymakers concerned with the fluctuating quality of our nation's teaching force.

We recommend that NCES make a similar resource commitment in the design of the proposed longitudinal study by having initial years of data collection staggered across multiple base years. In other words, we recommend that NCES *not* conduct this longitudinal study by following a single cohort of teachers teaching in only one base year, but rather that NCES use several base years and follow several initial cohorts of teachers over time. Our recommendation is similar to the strategy that NCES employed when their data collection program for secondary school students moved from the single cohort NLS-72 to the dual cohort HSB.

Recognizing the cost implications of this recommendation, we do not make it lightly. We believe strongly, however, that if NCES is to conduct this longitudinal study, they would be best served by doing it well, recognizing the rival hypotheses that will plague data analysts in the years ahead. The inclusion of multiple base years is, in our opinion, the best way that NCES can prevent researchers from being totally stymied by the "Age-Period-Cohort" problem (Fienberg & Mason, 1987). Below we give a brief overview of the problem; in Section 4, we delve into its implications in greater detail.

When a researcher describes a teacher's place in his or her career, the teacher's "place in time" can be measured by each of the following three markers: (a) his or her entry cohort (the year the teacher started teaching), (b) the year of his or her career (1st year, 2nr year, known more generally as "age" or "experience"), and (c) by the chronological year in which data are being described (1995, 1996, known more generally as "period"). All three time metrics can be substantively important. Teachers may have certain attitudes or go through certain transitions as a function of any one of them. Teachers may be affected by the particular year they entered teaching (the cohort effect), the particular year of their career

(the age effect), and by the particular year that is being described (the period effect).

Setting aside the obvious complications that arise for teachers who have had breaks in service, knowledge of any two of these time markers fully defines the third. If, for example, a researcher collects data on the 3rd year of the career for a teacher who entered in 1990, the chronological year being described must be 1993. Or if the chronological year of data collection is 1996, then data describing 1st year teachers will only describe those in the 1996 entry cohort and data describing 2nd year teachers will only describe those in the 1995 entry cohort. This confound makes it difficult to separate out whether teachers behave the way they do because of their entry cohort, year of the career, or the particular time period being referenced.

Research on teachers' career paths has been plagued by this problem. When Mark and Anderson (1985) published one of the early longitudinal studies of teachers' careers, they concluded that teachers hired in the mid-to-late 1970s were much less likely to stay in teaching than teachers hired in the same school districts before them. But in a reanalysis of these same data, Singer and Willett (1989) showed that this supposed "entry effect" was more likely attributable to a "period" effect--teachers hired in the later cohorts were subjected to RIFs in the early 1980s, thus giving them the appearance of having "less commitment" but actually their departure appeared to be a function of policies beyond their control.

Cross-sectional data on the teacher career totally confounds all three sources of information about time. When researchers analyze data from the 1990 SASS and detect differences across teachers with different years of experience, they cannot know whether these differences are attributable to the "age" of the teacher in his or her career, the particular entry cohort (are teachers hired in the 1980s less committed, for example, than those hired in the 1960s), or the effects particular to that individual chronological year (1990).

Longitudinal data collection represents a first step towards unraveling these effects. If NCES collects longitudinal data on teachers who have already amassed varying years of experience in the base year of data collection, researchers can attempt to separately identify some of the effects by, for example,

comparing the responses of teachers in their 1st year of service who were surveyed in the base year of the longitudinal study to those who were in their 1st year of service in a later year of longitudinal data collection.

But if all longitudinal data collection begins in the same base year, there will still be inseparable confounds (Baltes, 1972). Moreover, NCES would have data describing the crucial first few years on the job only for one entry cohort (as subsequent data collection would focus on later years of the career). We therefore recommend that NCES commit to beginning the study of the new longitudinal cohorts of teachers in at least two, and preferably three, base years. In Sections 4 and 6, we detail several possible alternative data collection schedules resulting from this recommendation.

### 3.6 Principle #6: NCES must collect data from at all relevant levels of the organizational hierarchy

Recognizing that teachers exist within complex organizations, NCES embedded the teacher interview component of the SASS in a series of linked surveys across several levels of the organizational hierarchy. Related questionnaires were sent to the LEA, the school principal, and in some cases, to another school administrator. But, to keep costs to a minimum, the TFS component of SASS only surveys teachers, thus providing no organizational linkage for this second data collection point.

We recommend that NCES reconsider this omission when the future longitudinal study is designed. We believe that it is a serious omission to isolate the teacher data collection effort from its organizational environment. With the current TFS, for example, we have no way of knowing to what type of district any "movers" have moved. Movers are asked a handful of questions, but it is difficult to determine whether this new district (or school) is different in fundamental ways from the teacher's previous district. Without such information it is difficult to build statistical models that might explain why the teachers left. (To see how difficult it is to build such model s on the basis of available data, consult the Choy, et al., 1992.) Moreover, if NCES would like to use other respondents as informants

about the target teachers, such multi-level data collection is imperative. By sending questionnaires to the principal of each sampled teacher's school, NCES might be able to collect a wealth of contextual data that might explicate the teacher's position in the school. Without linked questionnaires, researchers will have no way of corroborating the answers of the sampled teachers.

Not only do we recommend that NCES collect data on the organizational context of teachers as they move from school to school and district to district, we recommend that NCES also consider adding another level of data to the longitudinal study: data describing the *students* that the teachers serve. One way of collecting such data would be to include additional explicit items on the teacher questionnaire, items asking teachers to describe the students in their charge in greater detail than is presently done. Another suggestion that we make here, and recommend that NCES consider, is to build explicit linkages between the new longitudinal study and other on-going NAEP data collection programs that include students.

Realistically, such linkages can only be implemented for a subset of teachers. Teachers will move schools, school districts, and states immediately after the base year of data collection. For the truly longitudinal study to accurately describe teachers' careers, *all teachers will need to be followed as they move jurisdictions.* Indeed, teachers who leave teaching will need to be followed out of the profession so that NCES can collect data on them if and when they reenter the schools. But even amongst those who stay, we would expect a reasonable amount of mobility. Tracking these movers to their new schools will represent a data collection challenge. But if it is possible to link only some of these teachers (even as they move) to NAEP data collection, this linked study would undoubtedly prove a vital resource to researchers in the years ahead.

## 4.0 Identifying the target population: Whom should NCES study?

Identification of the target population for this longitudinal study is particularly complex because of the variety of paths that teachers may take through the schools. Each year after initial hire, a teacher decides, either consciously or by default, to exercise one of several career options--to stay in the same school, to move to another school (which may be in the same district, or not), to teach the same subject matter and grade level (or not), or to leave teaching. The current teaching force in any given year is therefore composed of a heterogeneous group of individuals who vary with respect to the year that they began teaching, the number and duration of career interruptions, and the total years of teaching experience. To select an appropriate target population for study in the proposed new study, we must first consider the sources of career heterogeneity in the teaching force by unpacking the diversity of teachers' career paths.

### 4.1 Diagramming the teaching career

Begin by conceptualizing the teaching career as a three state stochastic process with annual transition points. The career begins when a teacher is hired to work in a particular school (**Entry**). At the beginning of the next school year, the teacher can occupy one of three states: (1) **Stay** in the same school; (2) **Move** to another school; or (3) **Leave** teaching. Teachers who stay in the same school or who move to another school face the same three options in the subsequent year. Teachers who leave teaching have only two options: they can **remain out of** teaching or they can **return** to teaching. These same two sets of options are available in each subsequent year. Teachers who return can occupy any of the three possible states that can be occupied by a current teacher whether or not the teacher had already left once--they can either stay, move, or leave. This last component of the process is critical because leaving

teaching has not turned out to be the absorbing state it was once thought to be. The phenomenon of returning teachers is now acknowledged as an important component of the teaching force, and the reserve pool of former teachers represents a vital source of teacher supply.

Analysts of the teaching career will find it necessary to distinguish between Movers and Stayers, but from a design perspective, it is possible to collapse these two states into one--**In teaching**. Doing so greatly simplifies the representation of the teaching career, and in this simplification, we may find a strategy for articulating a reasonable target population from which to sample. In this simpler representation, in any given year after entry, teachers can occupy one, and only one, of *two* states--**In teaching** and **Out of teaching**. Teachers who are "in teaching" may be teaching in the same school as they were in the previous year or they may be in a different school; they may be teaching the same subject and grade level or a different subject and grade level. Variation with respect to job assignment is ignored; all that is relevant in this representation is whether or not the teacher is currently teaching.

Figure 1 presents the possible paths that teachers in this two-state stochastic process may take as they move through four years of transitions. During these four years, teachers may follow one of $2^{(4-1)} = 8$ possible paths; the 8 paths can be further categorized into four general prototypes. Teachers following **path A** enter and stay in teaching uninterruptedly. Teachers following **path B** enter teaching and then leave, not to return within the time period.

Figure **1.**
The four prototypical paths through the teaching career

<table>
<tr><th>Hire:<br>Year 1</th><th>Year 2</th><th>Year 3</th><th>Year 4</th></tr>
<tr><td rowspan="8">Entry:<br>In teaching</td><td rowspan="3">In teaching</td><td rowspan="2">In teaching</td><td>A: In teaching</td></tr>
<tr><td>B: Out of teaching</td></tr>
<tr><td rowspan="2">Out of<br>teaching</td><td>C: In teaching</td></tr>
<tr><td rowspan="2"></td><td>B: Out of teaching</td></tr>
<tr><td rowspan="3">Out of<br>teaching</td><td rowspan="2">In teaching</td><td>C: In teaching</td></tr>
<tr><td>D: Out of teaching</td></tr>
<tr><td rowspan="2"></td><td rowspan="2">Out of<br>teaching</td><td>C: In teaching</td></tr>
<tr><td></td><td>B: Out of teaching</td></tr>
</table>

Teachers following **path C** enter teaching, leave at some point, and then return within the time period. Teachers following path **D** are those who left, returned and then left again.

As the years unfold, the number of possible distinct career paths increases exponentially. After 10 such years, each member of an entering cohort of teachers could have followed one of $2^{(10-1)} = 512$ distinct paths! Although some of these paths would be particularly common (such as teaching continuously for 10 years or teaching for only 1 year, never to return) and others would be rare (alternating in and out in each year), there are a dizzying array of options. But overarching this complexity is the simplicity of the four prototypical paths outlined in Figure 1: A--teaching continuously; B--taught continuously for some period of time, not currently teaching; C--taught, left, returned, (may have left again and returned) and still teaching; and, D--taught, left, returned (may have left again and returned) and not currently teaching.

Identification of these four prototypical career paths allows us to extend Figure 1 over time to create a more realistic and complete diagram of the teaching career. In particular, it allows us to document the many career paths taken by *all* teachers who were hired in any of the entry years manifested amongst the stock of teachers actually teaching in any given chronological year. To keep the presentation general, let the current year--which will be a base year for data collection in the longitudinal study--be designated as "Year X." Figure 2, which represents these many career paths was developed from Figure 1 by:

- **extending time,** having the time line continue for many more years beyond the fourth year after entry. We have illustrated this continuation by adding a wide shaded column representing the occupational states the teachers could have occupied in years 4, 5, 6, 7, ...., through the year immediately preceding "year X," the base year of data collection.
- **adding entry cohorts,** by having replicate tables for each of the possible entry cohorts. To be complete, we would have to go back as many entry cohorts to pick up the entry year of *every current teacher* in the base year of data collection, year X. For simplicity, and without loss of

generality, we present only three entry cohorts--an "early," "middle," and "recent" cohort.

Notice that the three tables within the figure are similarly structured, the only distinction being the length of the time period designated as "the intervening years." Were we to create the full set of tables necessary to depict every possible entry cohort needed to encompass all teachers teaching in the base year X, we would have many more tables. Were we to specify the full set of tables necessary for describing the teaching force of 1995, for example, we would have as many as 50 tables (going from 1945 to 1995), each with (1995 - entry year + 1) columns. Nevertheless, each table would have the same general structure.

First consider the difficulties in identifying a suitable target population in any single base year. Figure 2 highlights two major problems inherent in identifying such a target population. The first problem is the great *diversity* of the current teaching stock. Were NCES to go into the schools and compile a list of all current teachers, only those individuals "still teaching" would be included (as indicated by the shaded cells in the last column of Figure 2). This stock is comprised of two groups of teachers: (1) those who have *never had a career interruption* (those following path A); and (2) those who have had one or more career interruptions, but who are currently teaching (those following path C). A sample of teachers selected from this stock is therefore extraordarily heterogeneous with respect to career trajectories. Those following path A (those with uninterrupted careers) will come from a wide

Figure **2**: Understanding who is included and excluded in the current stock of teachers

| Hire: Year 1 | Year 2 | Year 3 | intervening years | Year X |
|---|---|---|---|---|
| Early entry cohort **In teaching** | **In** | **In** | as many years as necessary to get the teachers to year X, the base year of data collection for the longitudinal study | A: **In** |
| | | Out | | B: Out |
| | | | | C: **In** |
| | Out | **In** | | B: Out |
| | | | | C: **In** |
| | | Out | | D: Out |
| | | | | C: **In** |
| | | | | B: Out |

| Hire: Year 1 | Year 2 | Year 3 | intervening years | Year X |
|---|---|---|---|---|
| Middle Entry cohort: **In teaching** | **In** | **In** | as many years as necessary to get the teachers to year X, the base year of data collection for the longitudinal study | A: **In** |
| | | Out | | B: Out |
| | | | | C: **In** |
| | Out | **In** | | B: Out |
| | | | | C: **In** |
| | | Out | | D: Out |
| | | | | C: **In** |
| | | | | B: Out |

| Hire: Year 1 | Year 2 | Year 3 | intervening years | Year X |
|---|---|---|---|---|
| Recent Entry Cohort: **In teaching** | **In** | **In** | as many years as necessary to get the teachers to year X, the base year of data collection for the longitudinal study | A: **In** |
| | | Out | | B: Out |
| | | | | C: **In** |
| | Out | **In** | | B: Out |
| | | | | C: **In** |
| | | Out | | D: Out |
| | | | | C: **In** |
| | | | | B: Out |

variety of entry cohorts. And those following path C will not only come from a wide variety of entry cohorts, they will also vary widely with respect to the total *number of different states* they have occupied (how many times they have left and returned) as well as with respect to the *duration* in each of those various states.

Some might argue that this heterogeneity is a problem of analysis, not design. If NCES measures these attributes (entry cohort, number of different states, and duration in each of the various states), data analysts could use this information as covariates in statistical analyses. But because teachers will not be equally distributed with respect to these career attributes (as we shall show in section 4.2), allowing these sources of heterogeneity to vary naturally in a sample may render the base year sample for the longitudinal study too heterogenous for detailed analysis. A simple look at the difficulty researchers have had in using the one year TFS to model attrition highlights the problems created by such heterogeneity (see, e.g., Choy et al., 1993). Thus, we believe it is important to consider the heterogeneity of the teaching stock with respect to **entry year, number of previous spells**, and **duration of previous spells** when identifying the target population for study.

The second lesson that we can learn from Figure 2 concerns the potential bias associated with using the entire teaching stock as the target population. The teaching stock in any given year is comprised of two groups of teachers: (a) those following the uninterrupted career path A; and (b) those following the interrupted career path C who happened to be back in the schools in that year. A sample drawn from this stock cannot be used to generalize back to any particular entry cohort because it omits the colleagues of these eligible teachers who started at the same time (in the identical entry cohort) but who left teaching at some point and who are not currently teaching (all those in paths B and D). Rather, such a sample would generalize only to the stock of teachers who happened to be teaching in a that common base year. Hoem (1985), following Ryder (1965), classifies this bias as "selection by virtue of survival."

Were the "teaching force" an enclosed system (like a sink with a faucet pouring water in at a

constant rate and a drain allowing water out at a potentially different, but nevertheless constant rate), the effects of the omission of peers following paths B and D could be estimated on the basis of concomitant information. But due to well-established perturbations in the system arising from: (a) variation across chronological years in the size of the entering cohorts; (b) variation across chronological years in the size of the non-voluntary exiting cohorts; and (c) variation in the risk of moving schools and leaving teaching as a function of entering cohorts, it is virtually impossible to estimate the size of the potential selection bias.

## 4.2 What *do* we know about the stock of current teachers?

Although the stock of teachers teaching in any particular year is the end result of a complex interplay of too many ill-understood factors to be modeled fully on the basis of available data, we do have some information from the SASS and TFS that can be used to make ballpark estimates as to the relative composition of the teaching stock in a future base year for the planned longitudinal study. To help understand how the factors identified above might manifest themselves in an actual longitudinal study, we now briefly present data from the 1987-88 SASS and the 1988-89 TFS that provide some information about the distribution of number of spells and years of service in each of those spells. Because these data are retrospective, they suffer from the very problems and biases outlined in section 3. However, as long as we remain cognizant of these shortcomings, we believe that these cross-sectional results can provide some useful information for study design.[3]

**4.2.1 Number of spells.** In outlining the possible career paths in Figures 1 and 2, we assumed that some teachers who left the schools would eventually return to teaching. This presumption was based upon findings from longitudinal studies of beginning teachers conducted at the state level (see, e.g.,

---

③ Note that we have chosen to use the SASS-I and TFS-I because at the time of writing, the TFS for SASS-II was not yet available. Although the actual point estimates will obviously differ between the two sets of surveys, we do not expect the general conclusions to differ widely.

Murnane et al, 1991; Willett & Singer, in press) in which researchers have estimated that approximately one-quarter to one-third of all teachers who leave teaching will eventually return. Although asking *current* teachers whether they have ever had a career interruption cannot tell us about teachers who left and never returned (thereby precluding estimates of these percentages on a national basis), such questions can describe those teachers who follow paths A and C.

Towards this end, we examined teachers' responses to the question about whether they had ever had "breaks in service of one year or more" since they began teaching. We used this information to determine whether teachers were in their first teaching spell, second teaching spell, etc, where "spell" is defined as years of continuous service broken by an absence of one or more years.

Reported breaks in service, although common, are by no means the norm. Approximately two-thirds (65.4%) of the teachers reported that they were in their first teaching spell, another one-quarter (23.8%) reported that they were in their second teaching spell, and only 10.8% reported that they were in their third spell (or higher). Of course, the distribution of spell number varies by entry cohort. As shown in Figure 3, among teachers hired in the 1980s, less than 15% reported having a break in service whereas among teachers hired before 1965, more than half reported having at least one such interruption. Note, however, that at least part of this relationship is due to the varying lengths of time in which these different groups of teachers *could* leave and return. Were we to have longitudinal records on teachers in the later entry cohorts that were as long as those we were able to reconstruct for teachers in the earlier cohorts, we would likely find that many of them would also ultimately leave and return as well.

This information on the distribution of the number of previous spells amongst the current stock of teachers has at least two implications for the design of a longitudinal study. First, it documents the heterogeneity of the stock of current teachers with respect to reported spell number. Because approximately one-third of the teaching stock will have reported at least one break in service, it becomes crucial that NCES either collect extensive data describing these teachers' entire career histories or that these teachers be set aside from the target population for the longitudinal study. Without understanding

the earlier years of these teachers' careers, it will be difficult to situate their data at the appropriate place on the longitudinal time axis. Second, it documents the strong relationship between entry cohort and spell number. In any stock sample, it is the teachers in the oldest entry cohorts who are most likely to have reported breaks in service. Selecting a target population using information on breaks in service will necessarily affect the distribution of entry cohort within the target population.

**4.2.2 Years of service during each spell.** We can also use information from the SASS to estimate the distribution of years of continuous service within each spell, as in Figure 4. Although we have drawn the figure so that the percentages within each spell sum to 100 (thereby facilitating comparison of the distribution of spell *length* across spells) recall that approximately two-thirds of the teachers are in their first spell (the most forward density in Figure 4), that nearly one-quarter are in their second spell (the middle density), and that less than one tenth are in their third spell or more. This means that were we to collapse *across spells* and examine the distribution of length of time in spell, this distribution would most closely resemble that for teachers in their first spell.

Begin with teachers in their first spell. Note that it is only for this group of teachers that the distribution of *spell length* is identical to the distribution of years of teaching experience. Examination of the first density in Figure 4 suggests that we can divide the experience axis into three broad groups: those in early career--years 1 through 5--which accounts for approximately one quarter (25.8%) of these teachers; those in mid-career--years 6 through 19--which account for approximately half (51.5%) of these teachers; and those is later career--years 20 and more--which accounts for the remaining quarter (27.2%) of all first spell teachers. The median years of full-time experience for those in their first spell is 12 years.

The distribution of spell length for teachers who have had one, or more than one, break in service (the remaining two densities in Figure 4) is broadly similar, but is skewed more towards shorter lengths. Although the early years--first through fifth--account for only one quarter of the first spell teachers, they account for approximately one-third (34.6%) of the teachers who have had only one break in service and

Figure 3

# *Distribution of number of spells by entry cohort in 1988 SASS*

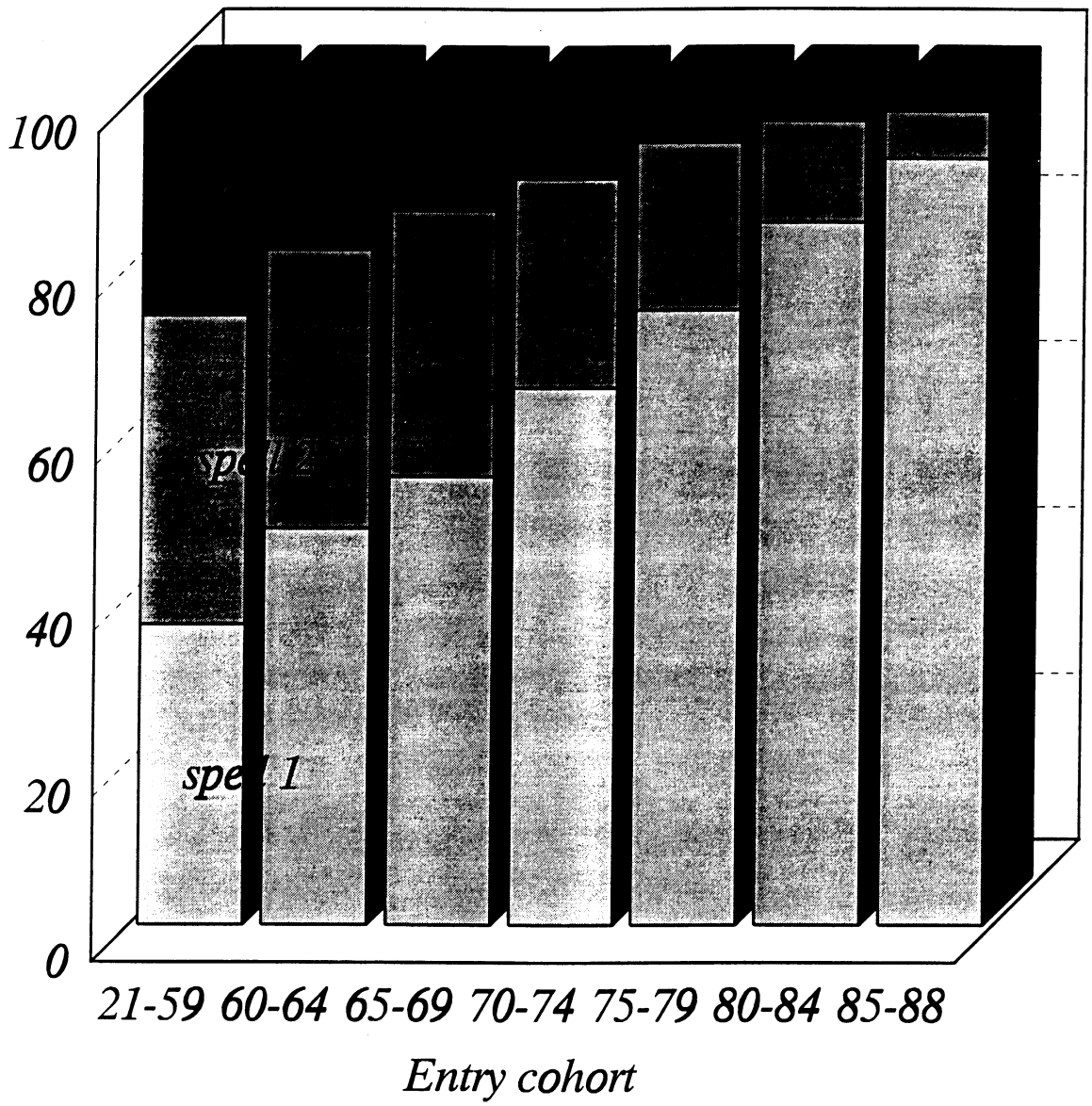
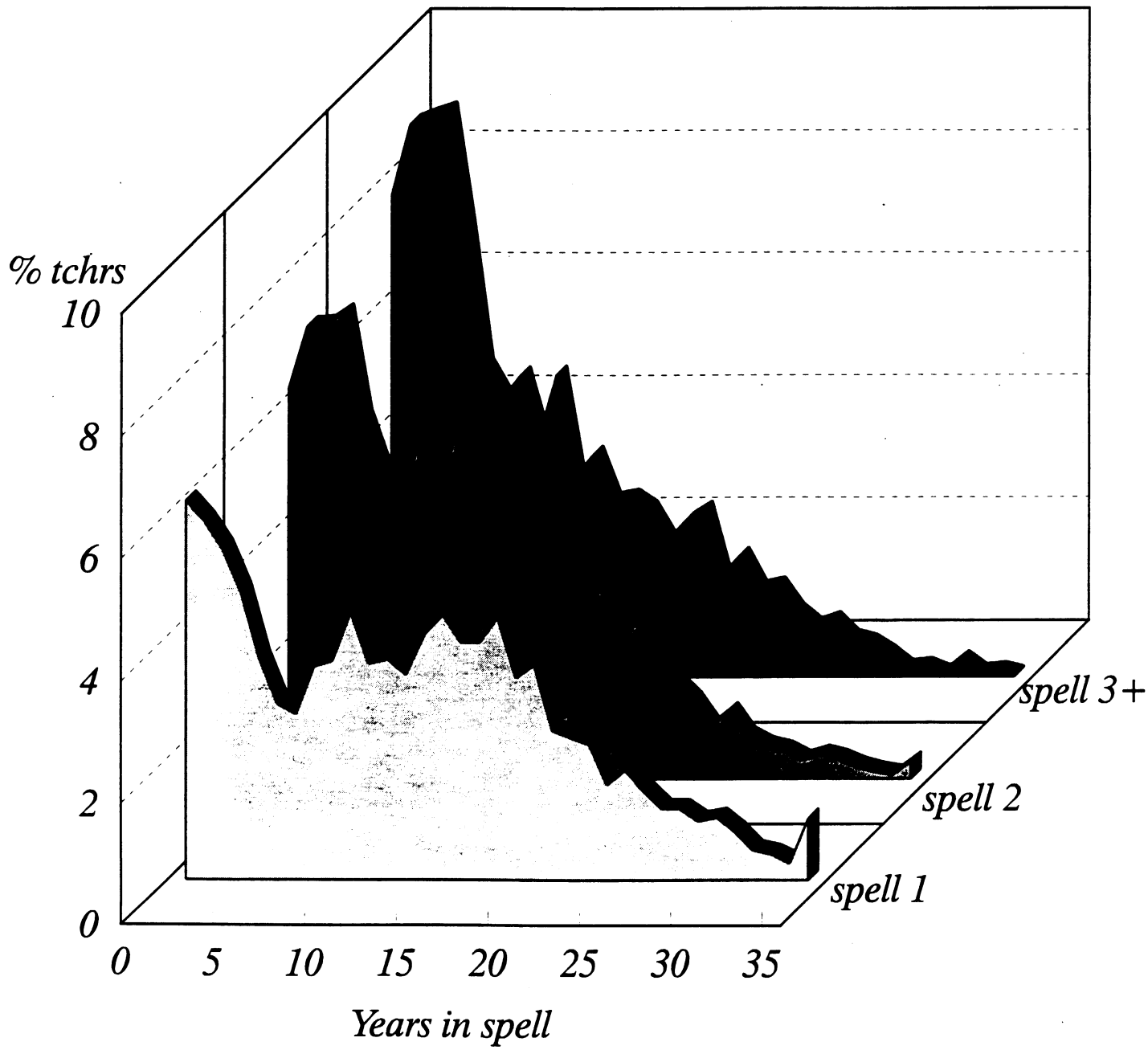


Figure 4

# *Years in spell by spell number*

## *1988 SASS*



42.5% of the teachers who have had more than one break in service. As a result, the median spell length for these two other groups is 9 years and 7 years respectively. To be sure, part of this relationship between spell length and spell number is attributable directly to the nature of retrospective data in a cross-sectional study. Because all teachers were interviewed in a common year, many of those teachers in the later spells will not have yet had the opportunity to stay for long periods of time.

What implications does this information have for the design of the longitudinal study? First, it documents the preponderance of teachers in the early years of their careers. Were we to select a focused target population based upon the principle of trying to include as large a fraction of the teaching force as possible, we would focus primarily on teachers in early-, and perhaps mid-, career. Second, by documenting the relationship between years in spell and spell number, it shows that were we to stratify teachers by year in spell and perhaps focusing on teachers in the early years in their spells, we would inevitably include large numbers of teachers who have already had a career interruption.

### 4.3 Four possible target populations

Recognizing the heterogeneity of the teaching stock with respect to the number of career interruptions, the duration of in-teaching spells, and the year of entry into teaching, let us now consider four possible ways of defining the target population for longitudinal followup. We will discuss the related issues of length and periodicity of data collection in section 5.

Figure 5 presents four alternative target populations: (a) a stock sample; (b) first year teachers; (c) beginning teachers; and (d) a stratified stock sample. These four--each having advantages and disadvantages--are not the only possible designs, and NCES may wish to combine elements of each or further restrict their definitions. Coupled with the continued pairing of a cross-sectional SASS and a one-year TFS conducted at periodic intervals (as discussed later), each of these longitudinal studies would provide a powerful basis for learning about *some aspect of the teaching career.* The conundrum for design is that each is optimally suited to answer a different set of questions.

**4.3.1 Stock sample of teachers.** This design, what most observers would consider the obvious choice, was the very design proposed by NCES during the January 1993 planning conference. It has many advantages, including: (a) the relative ease with which the universe can be listed; (b) the coverage of teachers with a wide array of career histories (because all teachers have a non-zero probability of selection); (c) the ease with which subgroups--e.g., beginning teachers, first year teachers, teachers of certain subject specialties, teachers in "high risk" geographic areas--could be oversampled; and (d) the clear correspondence with a base year SASS.

But use of a stock sample of teachers has at least three major disadvantages. First, by choosing to include *all* current teachers, the study will be decidedly unfocused. Were sample size and money not at issue, this lack of focus might not be problematic. Relevant groups for

Figure 5 Four possible target populations

| **Full stock** | 1st year | 2nd year | 3rd year | 4th year | 5th year | ... | nth year |
|---|---|---|---|---|---|---|---|
| 1st spell | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| 2nd spell | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| ... | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| nth spell | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |

| **First year teachers** | 1st year | 2nd year | 3rd year | 4th year | 5th year | ... | nth year |
|---|---|---|---|---|---|---|---|
| 1st spell | ✔ | | | | | | |
| 2nd spell | ✔ | | | | | | |
| ... | ✔ | | | | | | |
| nth spell | ✔ | | | | | | |

| **Beginning teachers** | 1st year | 2nd year | 3rd year | 4th year | 5th year | ... | nth year |
|---|---|---|---|---|---|---|---|
| 1st spell | ✔ | ✔ | ✔ | ✔ | | | |
| 2nd spell | ✔ | ✔ | ✔ | ✔ | | | |
| ... | ✔ | ✔ | ✔ | ✔ | | | |
| nth spell | ✔ | ✔ | ✔ | ✔ | | | |

| **Stratified stock** | 1st year | 2nd year | 3rd year | 4th year | 5th year | ... | nth year |
|---|---|---|---|---|---|---|---|
| 1st spell | ✔ | | ✔ | | ✔ | | ✔ |
| 2nd spell | ✔ | | ✔ | | ✔ | | ✔ |
| ... | ✔ | | ✔ | | ✔ | | ✔ |
| nth spell | ✔ | | ✔ | | ✔ | | ✔ |

oversampling could be identified, and NCES could be sure to include sufficient numbers of teachers in each of the important "oversample" cells to enable separate statistical analyses for all relevant subgroups. But sample size and cost must be considered, and in this regard, a stock sample with the necessary oversamples would likely be unfeasible. Consider, for example, that in the one-year TFS, which includes nearly 7,500 teachers with dramatic oversamples of leavers and movers, it has proven difficult to identify factors associated with teachers' stay, move, or leave decisions. There is no reason to believe that such difficulties would not persist.

The second concern with use of a stock sample of current teachers is that they are too heterogeneous with respect to the career paths they have already taken. As we have outlined in the previous sections, a stock sample of teachers--even with an oversample to ensure adequate representation of specific subgroups--will vary with respect to entry cohort, number and duration of career interruptions, and length of time in current spell. Although NCES could attempt to retrospectively reconstruct every teachers' event history from licensure to the base year of data collection, for many of these teachers, this process would involve retrospecting back many, many years. Examination of data from SASS I provides clear evidence of rounding errors (the tendency, for example, to report total years of experience using numbers ending in 0 and 5). Rounding errors, as well as memory lapses and telescoping would likely prove problematic when attempting to retrospectively reconstruct career histories for all sampled teachers.

The third concern with the use of a stock sample is bias. As we discussed earlier in this section, a stock sample only includes teachers who followed paths A and C in Figure 2. Colleagues of these teachers who happened to not be teaching in the particular base year of data collection (but who might have returned to teaching in the following year--those on path D) would not be included. This omission would make it virtually impossible for researchers to use data from such a sample to correctly estimate hazard and survivor functions documenting teachers' career paths and to correctly evaluate the effects of early career experiences on teachers later in their careers.

How problematic is this selection bias? For external opinions, we present the views of two methodologists experienced in studying employment histories and demographic trends. In describing the assumptions necessary to use a stock sample to model unemployment data, Lancaster (1990) describes what is known as the "initial conditions problem," where the researcher does not have access to the "initial conditions" of each member of the stock sample. He writes:

> In general, a fully satisfactory solution to the initial conditions problem appears to require a full model for the rate at which people enter the state in question. That is, we require to embed the stock sampled data in a stochastic process describing the full state biography of each individual. (p. 189)

The problem with a longitudinal study of teachers, of course, is that the complexity of paths outlined in Figure 2 suggests that any researcher attempting to develop the necessary "full model" would be treading on dangerous ground. The alternatives that Lancaster offers are similarly problematic: (1) to ignore previous started spells and base all models only on those spells that begin after the sampling date; or (2) to assume stationarity in the stochastic process over time (e.g., to assume a constant flow into the schools over time). The lasting impression that Lancaster gives is that a stock sample is to be avoided if at all possible.

Hoem (1985) raises similar concerns, going even further to discuss how commonly tried remedies will fail to eliminate the bias. He writes:

> Selection by virtue of survival persists in the face of any attempt at drawing a fair (noninformative) sample from the target population as of the sampling date or of increasing the sample size, even to an exhausitve enumeration, and no vigor in the fight against nonresponse can drive it out of existence. It is present from the outset, built into the whole procedure, simply because it is impossible to interview the dead and possibly those who have out-migrated (p. 262).

Although Hoem's comments were directed towards the study of mortality, not employment, the problem of selection by virtue of survival is identical, the difference arising solely with respect to the particular

states that an individual can occupy. The unavailability of data describing teachers following paths B and D in Figure 2 renders a stock sample truly problematic for modeling teachers transitions through their careers.

**4.3.2 First year teacher study.** This design represents the opposite end of the design continuum. Recognizing that a stock sample would spread resources thinly over a heterogeneous group of people, and that lack of information about previous history might render the information from all but the newly hired teachers useful, the first year teacher study retreats from the notion of a broad target population and argues that NCES focus intensely on teachers in their first year of their current spell. The first year of the career has been identified by both qualitative and quantitative studies as the most crucial period in the teaching career (Johnson, 1991; Lortie, 1975; Murnane et al., 1991). By focusing on first year teachers and following them longitudinally (for a period of time to be discussed in section 5), this study would be sure to gather complete data describing all the sampled teachers during this crucial time period as well as comparable data for each successive year in these teachers' careers. Previous employment history could be captured by covariates, but the key idea here is that *the time clock for all sampled teachers would be starting at an identical moment.* Elsewhere (Willett & Singer, 1991) we have argued that the study of teachers' careers must track teachers from their entry into the profession to their exit (and beyond). A longitudinal study beginning in their first year on the job would offer researchers the ideal opportunity to track teachers through the profession.

A first year teacher study, as we conceptualize it, however, does differ from a classic study of an entering cohort of teachers in one prominent aspect--the first year teachers need not be in their first spell of teaching. We are using the term "first year teachers" here also to include those who are returning to the schools after a career interruption. Based on the SASS I, we estimate that 63% of first year teachers will be in their first spell, 23% will be in their second spell and 13% will be in a third spell or higher. We are not recommending that returning teachers be oversampled, but rather that they be sampled in proportion to their representation in the population of first year teachers that exists during the base year

of data collection.

We are including teachers returning after a career interruption in our specification of this target population because we believe that there is much new knowledge to be gained by comparing returning teachers to newly hired teachers to see whether they experience their careers in similar or different ways. In one of the only longitudinal studies of returning teachers (Willett and Singer, in press), we found that the first year back in teaching for returnees is at least as risky (if not more so) than it is for totally inexperienced first year teachers. Given that over half of the newly hired teachers in many school districts are, in fact, returning teachers, this group constitutes a pocket of serious policy importance. Of course, great care would need to be taken in constructing an interview to gather retrospective career information from those teachers not in their first spells. But as we show in section 6, such data collection could be conducted, and the data could then serve as covariates in subsequent statistical analyses.

A study of first year teachers is also not without its shortcomings. Perhaps the major concern is its specificity. Teachers in their first year of a spell comprise only 6.4% of the nation's teaching force. Although specificity is desirable, so is generalizability. It may be too draconian to limit the target population to only teachers in their first year. After all, any particular base year for the longitudinal study (be it 1988, 1999, or 2000) is unlikely to have policy significance. Thus, this longitudinal study would be following a single entry cohort that would be less and less policy relevant as it aged. Yet this criticism could be leveled at virtually every longitudinal study that NCES has conducted. There was nothing about 1972 that made it the "ideal" year to initiate a study of high school graduates (as in the NLS-72), nor was there anything particularly salient about 1980 for the initiation of High School and Beyond, nor 1988 for the initiation of NELS. The base year for a longitudinal study is often selected simply for convenience. What has to be relevant in the design of those studies and what is relevant in the design of this longitudinal study is that there be sufficient focus in the target population so that the results will generate useful information for researchers and policymakers.

A second shortcoming of a study of first year teachers is that it omits colleagues of returning

teachers who decided not to have a career interruption in the first place or who decided not to return to teaching by the particular base year of data collection. In other words, this design, unlike the stock sample design, does not include data on the colleagues of the returnees who entered in the same original years as these second and third spell teachers. This omission would deprive researchers of an important and relevant comparison group, enabling researchers to investigate, for example, whether returning teachers more closely resemble their peers who never left or newly minted teachers.

A third shortcoming of a study of first year teachers is that it would take many years before it could yield any information about teachers in mid to late career. Even if this study were fielded for at least up to 10 to 12 years (as we recommend in section 5), it would provide only limited information about teachers worklives during mid-career (as they stabilize their roles in the schools) or late-career (as they consider retirement). Examination of data describing teacher transitions suggest that mid-career is a time of few changes and that the changes in late career are far more predictable, driven by retirement. But there remain researchers interested in the career decisions of teachers in late career, when individuals are often at the highest levels of the salary ladder, drawing considerable resources from their school districts.

**4.3.3 Two other options: A beginning teacher study and a stratified stock sample.** These two designs fall between the extremes of the first year teacher study and the use of the stock sample that we have described above. Unfortunately, in attempting to address shortcomings of one of these two core designs, each then inevitably acquires the shortcomings of the other design. The **beginning teacher sample** moves from the specificity of a first year teacher study by expanding the target population to include all beginning teachers, say those in their first 5 years of their current spell. The **stratified stock sample** moves from the great heterogeneity of the full stock sample by choosing teachers in specific years of their current spells for study. In each of these latter two designs, we would still be limited to those teachers currently teaching in the particular base year of data collection. Colleagues of these teachers who entered in the same year but who were not teaching in that base year would be omitted

(whether or not they would later return) making it impossible to generalize to any specific entry cohort (other than that of the particular base year using the subsample of first year teachers).

The clear advantage of the beginning teacher sample over the first year teacher sample is the improved generalizability. As shown earlier in Figure 4, we estimate that this target population includes approximately 30% of the current teaching force, and this percentage could be expanded somewhat simply by altering the particular cutoff year for data collection (moving, say, to 6 years). As with the first year teacher sample, those in this target population would vary with respect to spell, making it imperative that retrospective data on previous jobs be gathered from all sampled teachers not in their first spell.

The major downside of expanding the definition of eligible teachers beyond those in the first year of service is that this sample will necessarily suffer from the same biases inherent in any study in which individuals are sampled based on survival (as described above). Not only would teachers vary with respect to length of service during the current spell (making it imperative that retrospective data concerning the previous years of service during this spell be gathered), only those teachers who have *survived* to this point in their careers will be included. By limiting this target population to teachers near the beginning of their current spell, it should be possible to gather reliable and valid data describing certain aspects of their recent employment histories with a minimum of telescoping and omission. Yet there would be little way of eliminating the serious selection bias.

The stratified stock sample would also have improved generalizability over the first year teacher study, but by selecting teachers who vary widely with respect to their year in the current spell, it would make it very difficult to reliably and validly retrospectively reconstruct equatable career histories for everyone in the sample. Teachers in their 20$^{th}$ year of service could not be expected to reliably describe very much about their teaching lives 20 years ago, let alone 10 years ago or even 5 years ago. Here, generalizability and some coverage of the entire teaching career would be traded for an ability to reconstruct entire career histories.

### 4.4 What is our recommendation?

No single target population will be optimal for addressing all research questions about changes in the teaching force over time. Use of a complete stock sample or a stock sample stratified by spell number and year in spell has the advantage of broad coverage, thereby giving some information to all interested researchers. Use of a first-year teacher sample or a beginning teacher sample has the advantage of being focused, thereby probably providing sufficient information for many researchers to delve into this crucial phase of the teaching career in great detail. A final decision must also consider two other time factors discussed in section 5--the length of the overall longitudinal data collection effort and the periodicity of data collection. Nevertheless, it is still possible to articulate some strategies that can be used to make this difficult decision.

For us, the core issue surrounds *which aspect of time is thought to be most relevant for the study of teachers' careers.* We believe that it is the lack of knowledge about the relative importance of the many temporal aspects of teachers' careers that makes it impossible to argue exclusively for one, and only one, of these design choices. Developmental studies of adolocents provide a useful comparison. Most studies of school-aged children (including HSB, NELS, the LSAY) select students who are in a particular set of grades (e.g., tenth and twelfth; eighth and tenth; seventh, ninth and eleventh; first through fourth) and then follow these students longitudinally (on an annual or semi-annual basis) for a pre-selected period of time. To help unravel the age-period-cohort problem, the more sophisticated studies are replicated in multiple base years. This strategy is defensible because of the strong relationship between a child's chronological age and grade, the social, psychological, and educational salience of a child's grade in affecting child outcomes, and the developmental trajectories that we expect children to follow. Birth cohort effects and period effects are assumed to be relatively small in comparison to age (grade) effects.

When it comes to the teaching career, however, reasonable people may disagree about which aspects of time are the most salient. In addition, the possibility of breaks in service of varying lengths of

time further expands the ways of conceptualizing time beyond the classic three--entry year, year of the career, and chronological year--to include spell number, length of previous in-teaching spells, and length of previous out-of-teaching breaks. Should a teacher's "age"--number of years of service--cumulate over spells or should it be re-set each time the teacher re-enters teaching? Should a re-entering teacher acquire the entry cohort corresponding to the year of re-entry, or should she be treated as a member of her original entry cohort? Does "age" continue to remain salient to teachers after many years on the job? For students, there is a world of difference between tenth grade and eleventh grade. Is there a similar difference between the worklives of teachers in their tenth year on the job and the eleventh year on the job? If the construct "years of teaching" (or "years in spell" for that matter) ceases to be salient after several years of teaching, does it make sense to design a study that stratifies participants on the basis of a variable unlikely to be strongly associated with any of the important outcomes? Indeed, a more fundamental question can even be asked: Should "spell" be defined in terms of breaks in service or change of school?

These issues lead us to believe that a stratified stock sample--the design of choice for most longitudinal studies of school-aged children--is *not* the best choice for a longitudinal study of teachers. The research community simply lacks a sufficient corpus of evidence to suggest the continued salience of year of the career (or year of spell) as the primary way to place teachers' careers on a temporal axis. Not only would it prove difficult to identify teachers in specific years of service (e.g., the 1$^{st}$, 6$^{th}$, 11$^{th}$, and 16$^{th}$ years), we are not convinced that years of service, or years in spell, provide such a salient time metric that they be used to stratify the target population.

In some ways, the fallback position from this is to select a stock sample from the entire pool of teachers teaching in some specific base year. But given the interest in using these longitudinal data to model teachers' career decisions, the selection bias problems associated with this approach loom large. The longitudinal study of a stock sample would virtually preclude researchers from using these data to model teachers' careers. Although these longitudinal data will be gathered for purposes other than this

as well, we believe strongly that it would be foolhardy to design the study knowing, a priori, that its data would be ill-suited for answering some of the most pressing questions being raised at present.

These concerns lead us towards the two studies of beginning teachers--focusing either exclusively on first year teachers, or more broadly on teachers in the first few years in the spell. From a purely methodological perspective, the study of first year teachers is to be preferred because it eliminates any potential selection bias. From a policy perspective, however, the desire to have some data on teachers in mid-career may create a strong incentive to expand the target population to include some teachers in the first few years of the spell. The potential for selection bias persists here, but if NCES could use data from one of the already fielded SASS's or TFS's (in other words the 1987-88 or 1990-91 SASS's and their accompanying follow-up studies) to select teachers in the later years on the job, this bias could be estimated. In addition, if the study were designed to oversample teachers in their *first* year, a beginning teacher study could provide virtually all of the advantages of the first year teacher study with some acceleration along the years of experience access in the first few years on the job.

We recognize that either a first year teacher study or a beginning teacher study will not satisfy those researchers interested in studying teachers in mid- or late-career. But a single study cannot achieve all research goals, and we believe that there is sufficient indication that it is the *early* years on the job that are the most critical defining ones for educators entering the teaching profession. If interest persists in these issues, the longitudinal sample could be tracked for an even longer period of time. But, in our view, consideration of the research goals outlined in section 2 makes it clear that the kinds of changes and transitions of greatest interest are most likely to occur during the early years on the job. Moreover, the certain difficulties associated with using a stock sample of teachers as the target population for a longitudinal study suggest that NCES might waste much time and resources collecting data that would not enable researchers to address their research questions. A first year teacher study or a beginning teacher study does less, but can do it well.

Given this preference for a study of teachers early in their careers, let us now turn to issues of the

periodicity and duration of data collection under the assumption that the target population consists of teachers in the early years of their spells. Although the discussion that follows in the next two sections can easily be adapted if NCES decides to pursue a stock sample, we decided that our recommendations on the time dimensions of data collection could be made more specific, and hence more useful, if we adopted this particular focus.

## 5.0 The time dimensions of data collection: How often, and for how long, should data be collected?

We have argued earlier, under the second design principle, that there are two types of question about time (Willett & Singer, 1989). Design of a longitudinal investigation of the teaching career must conceive of time as both a "predictor" and as an "outcome," its specific treatment being determined by the nature of the particular research question being asked. When an investigator asks how a teacher attribute, such as professional commitment, changes over time, then commitment is treated as though it depends upon time. This means that, conceptually, commitment is the outcome and time the predictor. On the other hand, when an investigator asks how much time passes before a teacher experiences a critical event, such as leaving teaching, time is treated as the outcome and characteristics of the teacher's background, environment and treatment are the predictors.

To address each of these types of question--treating time as either a predictor or an outcome--a researcher must use a different statistical technique. The former requires methods developed for the measurement of change, such as *individual growth modeling* (Rogosa, Brandt & Zimowski, 1982; Bryk & Raudenbush, 1985; Willett, 1989); the latter requires methods developed for the analysis of the occurrence and timing of events, such as *survival analysis* (Singer & Willett, 1991, 1993; Willett & Singer, 1991, 1993). Investigators who will ultimately analyze data from this new longitudinal study of teachers' careers must be able to apply *both* techniques to the data and have their efforts result in parameter estimates with adequate precision and hypothesis tests with adequate statistical power. To achieve this goal, NCES must commit to collecting enough waves of longitudinal data on each teacher over a sufficiently extended period of time.

How many waves of data must be collected? Over what period of time must these waves be spaced? The answers to these questions are complex and there is no single answer that is uniformly

acceptable. Among other things, the answers depend upon the nature of the attribute that is changing and the events that are expected to occur. In this section, we address these twin issues by separately contemplating each of the two research questions about time: (a) first, conceiving of time as a predictor and considering the precision and reliability with which individual change can be measured provides a ready yardstick by which we can decide upon a minimum number of waves of data that ought to be collected; and (b) second, conceiving of time as an outcome and considering the statistical power of survival analyses designed to examine the occurrence of critical events in the teacher's life provides a yardstick by which we can determine how long prospective data collection must endure.

## 5.1 How many waves of data are required?

Two waves of data separated by one school year--as in the case of the SASS and the companion TFS--do not permit researchers to portray the complex patterns of growth that we expect teachers to display over their professional lifetimes. Nor do two waves of data permit individual change over time to be measured validly, precisely and reliably. If change is to be measured well, lengthy longitudinal data must be collected.

To determine exactly how many waves of data are required, we must consider three important issues: (a) the *shape* of the trajectories that we anticipate the individual growth will follow; (b) the *precision* with which we would like to measure the individual growth parameters that characterize these trajectories; and (c) the *reliability* with which we would like to distinguish among teachers on the basis of their individual changes. Each of these issues must be addressed for each of the constructs that will be measured in the prospective study of teachers' careers, leading hopefully to a design consensus across all constructs. Below, we separately comment on these three linked issues.

**5.1.1 The shape of the individual growth trajectory.** Among the many problems with the two-wave design is its failure to provide sufficient data to characterize the *shape* of each teacher's growth trajectory. Two waves of data tell researchers only about each teacher's status at two points in time;

there is no information about *how* the teachers got from the first time to the second. In that single year, perhaps all of the change occurred immediately after the beginning of the year? Or, perhaps the teacher changed at a steady rate, gaining equal amounts per month? Perhaps the rate at which teachers changed declined steadily throughout the year, with status leveling off as spring approached? What will happen as these teachers continue to teach? With only two waves of data separated by only one school year, it is impossible to know.

If a particular teacher attribute is changing over time in a complex and interesting way, the researcher can construct a reasonable and valid portrait of that trajectory only when sufficient waves of data have been collected on the attribute. The number of waves of data required to adequately characterize individual change in any particular attribute is directly related to the complexity of the mathematical model that describes the value of the attribute as a function of time. If change in the attribute follows a straight-line trajectory for each person in the population, for example, individual growth in the attribute can be described by a linear function of time. The *individual growth model* representing each individual's change then contains a pair of *individual growth parameters*--the *intercept* and *slope* parameters--that represent the attribute's initial value and rate of change for each person, respectively.

Under the individual growth modeling framework, we regard any heterogenity in individual change that exists across persons to be a result of inter-individual differences in the individual growth parameters, not a result of variation across individuals in the mathematical form of the individual growth model itself. In the case of linear growth, for instance, between-person heterogeneity in change is attributed solely variation in the individual intercept and slope parameters across people.

Conceptually, one may regard the individual growth model as a *within-person regression model* that represents the individual's change over time. During an analysis of change over time, regardless of the sophistication of the statistical estimation technique involved, the investigator essentially fits an appropriate individual growth model to each person's empirical growth record which contains, hopefully,

sufficient waves of data for the estimation to be completed successfully. As is the case in any regression analysis, if the model is to be fit and the parameters and their standard errors are to be estimated, the number of available datapoints must exceed the number of parameters being estimated, at least by one. In other words, where the two-parameter straight-line growth model is involved, no less than three waves of data are required. If individual growth in the attribute is more complex, minimal data requirements increase. Non-linear growth models require additional waves of data--quadratic models a minimum of four waves, cubic models a minimum of five.

The consequence of this relationship is that before we can decide on the number of waves of data collection, we need to know more about the shape of typical individual growth trajectories in each of the four domains of measurement (teacher quality, teacher work context, teacher worklives, and teacher career paths). Such an evaluation could use historical information drawn from the substantive literatures in each of the four domains, be based on expert testimony, or could be based on the collection of longitudinal pilot data for a small representative sample of teachers. Unfortunately, the existing SASS and TFS datasets are little help here because they contain an insufficient number of waves of data to provide any insight the shape of the anticipated growth trajectories.

A further complication is that there is no apriori reason to believe that individual growth trajectories will be described by identical mathematical functions in all domains of measurement; different constructs may require different growth models. Individual growth may be linear in one domain, exponential in a second, and cubic in a third. In general, fewer waves of data will be required in those domains in which individual growth has a less complex functional form. A future study of teachers must take these potential differences in growth trajectories across domain into account. Fortunately, though, designing a study so that change can be measured adequately in the domain in which individual growth has the most complex trajectory automatically provides a sufficient number of datapoints for the measurement of change in domains in which individual growth has a less complex trajectory.

**5.1.2 The precision of estimates of growth.** The preceding discussion provides a way of deriving minimum data requirements. Where growth is linear, three waves of longitudinal data ensures that individual intercepts and slopes can be estimated, with one degree of freedom left over to estimate model goodness-of-fit (including residual sums-of-squares and standard errors). But the existence of three waves of data does not ensure that the individual intercepts and slopes are estimated *well*. Just as the precision with which any regression parameter can be estimated improves as further datapoints are added to the sample, so will the precision with which the individual growth parameters are estimated be dramatically improved as additional waves of data are included in a longitudinal design.

The specific mathematical form of the relationship between the precision with which the individual growth parameters can be estimated and the number of waves of data also depends upon the shape of the trajectory used to model individual growth. We illustrate the relationship between precision and waves of data in Figure 6, assuming that individual change can be adequately represented by a linear growth model, that occasions of measurement are equally-spaced, and that an ordinary least-squares (OLS) estimate of linear slope is an appropriate summary of the individual rate of change.[4] In the figure, we display the standard error of the individual rate of change (in units of residual standard deviation) as a function of the number of waves of data collected. Individual growth that is more complex will necessarily require additional waves of data if requisite levels of precision are to be achieved.

Notice that for designs with more waves of data, the standard error associated with the estimated linear slope is considerably smaller, reflecting an improvement in the precision with which individual change is measured. All else being equal, when five waves of data are available (as opposed to three, the

---

④ Under the assumptions stated in the text, the standard error of the slope (in units of residual standard deviation) is given by:

$$\frac{s.e.(slope)}{s.d.(residual)} = \sqrt{\frac{12}{T(T^2-1)}}$$

where T is the number of waves of longitudinal data that have been collected.

minimum number required in the case of linear individual growth), the standard error of the slope is more than halved; when seven or eight waves of data are available, the standard error drops to about a quarter of its three-wave value. This indicates that the precision with which individual change can be measured increases dramatically with the addition of a couple of extra waves, doubling when five waves are collected and quadrupling with seven or eight. Although this example has been based on an assumption of linear individual growth, it provides a very strong argument for planning to collect more than the minimum number of waves of data when designing this new longitudinal study. Similar arguments could be made for the case of more complex, non-linear individual change.

**5.1.3 The reliability with which change can be measured.** When we use the precision with which the intra-individual change can be estimated to suggest an appropriate number of waves of data, we are focusing on the individual respondent and his or her particular history of change over time. By including a sufficient number of waves of data in the design, we ensure that the change over time of each and each and every person in the sample can be measured at an acceptable level of precision.[5] Social scientists, however, are often less interested in questions of intra-individual change (which we will henceforth refer to as "Level-1" questions) and more interested in questions of *systematic inter-individual differences in change* (or "Level-2" questions). At level-2 we are concerned with the relationship between individual change and selected background characteristics of the individual, and we ask questions such as: Does the professional efficacy of beginning teachers decline less rapidly for those who participate in a mentoring program (in comparison to teachers not in such a program)? This focus on inter-individual differences leads social scientists to be concerned with the *reliability* with which change can be measured--reliability in this context being a group-level parameter that describes the extent to which individuals can be distinguished from each other on the basis of their individual changes

---

⑤ Given that the within-person residual variance (an estimate of the measurement error variance for a *single* individual, see Willett, 1989) usually differs considerably from person to person across the sample, it may be that different numbers of waves of data will have to be gathered on each individual in order to ensure that a constant level of precision is attained across all individuals.

(see Rogosa, Brandt & Zimowski, 1982; Willett, 1988).[6]

As one might expect, any improvement in reliability that results from collecting greater numbers of waves of data is partly a natural consequence of the corresponding increase in precision with which individual change can be measured for such designs (as is displayed in Figure 6). When individual change is measured more accurately, we are better able to distinguish among individuals on the basis of these changes. However, by virtue of the definition of reliability as a group-level parameter, its magnitude is affected not only by the precision with which individual changes have been measured but also by the magnitude of the heterogeneity in true change in the target population. Target populations with zero heterogeneity in true change (i.e., in which all members are experiencing identical true change) make it impossible to distinguish among individuals on the basis of these changes and the reliability parameter is necessarily zero. As heterogeneity in true change increases in a target population, then so does reliability (for a constant level of measurement precision).

The critical disadvantage of the reliability parameter is that it confounds the effect of within-person measurement error variance with between-person heterogeneity in true change. When either measurement error variance is large or when heterogeneity in true change is small, reliability will tend to zero. When either measurement error variance is small or when heterogeneity in true change is large, reliability will tend to unity. For these reasons, it is difficult to interpret any particular value of the reliability parameter as separately indicating levels of either measurement error variability or variability in true change. This confounding must be taken into account when interpreting values of reliability.

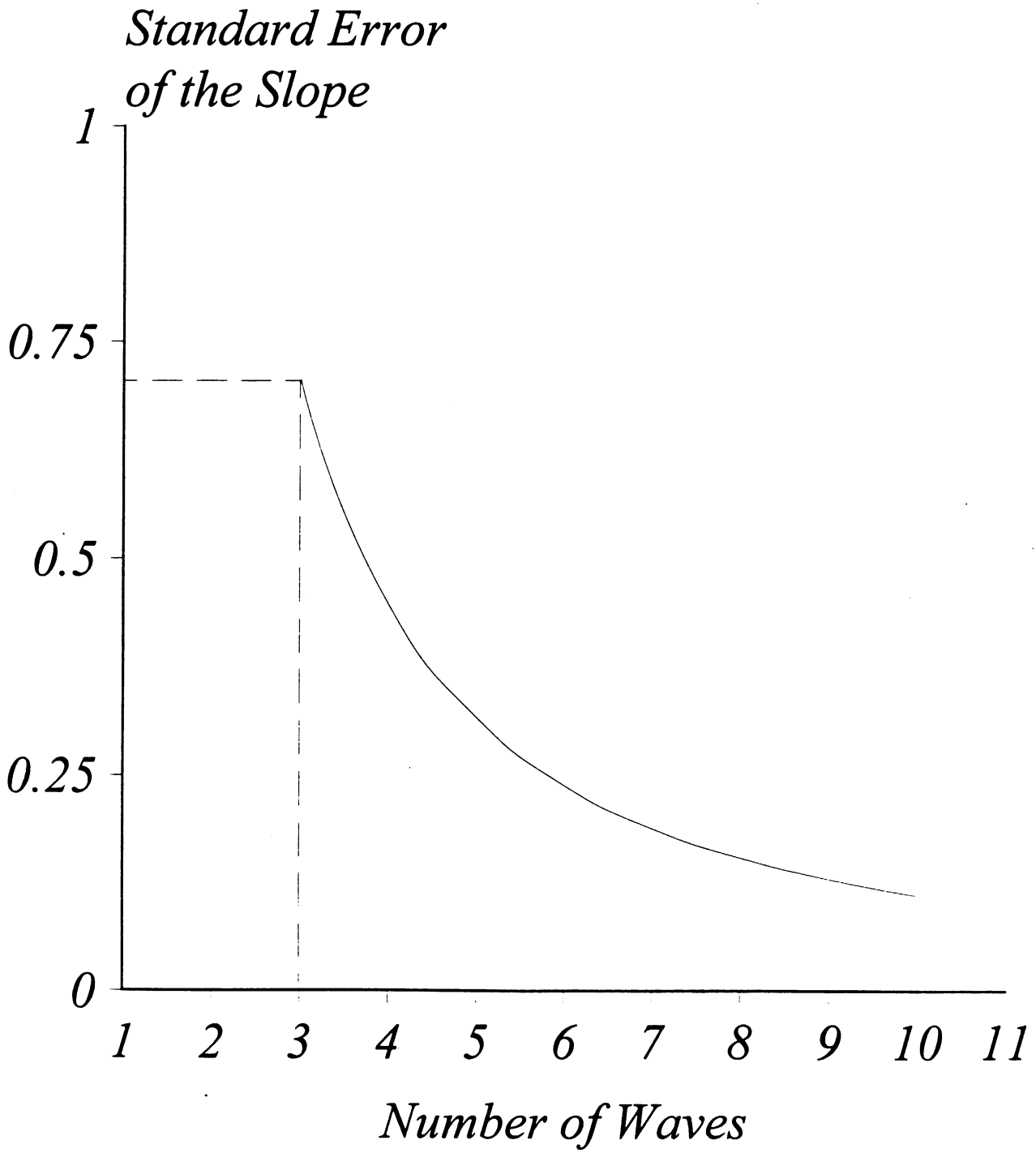
Nevertheless, under the same set of assumptions that underpin Figure 6, we can display the reliability of the linear rate of individual change as a function of the number of waves of data collected. In this case, for reasons alluded to above, the relationship between the reliability of change and the number of waves of data also depends on the variance of the true linear slopes across individuals in the

---

® The reliability of change is defined as the proportion of the population variance in observed change that is variance in true change (see Rogosa, Brandt, & Zimowski, 1982; Willett, 1988).

Figure 6

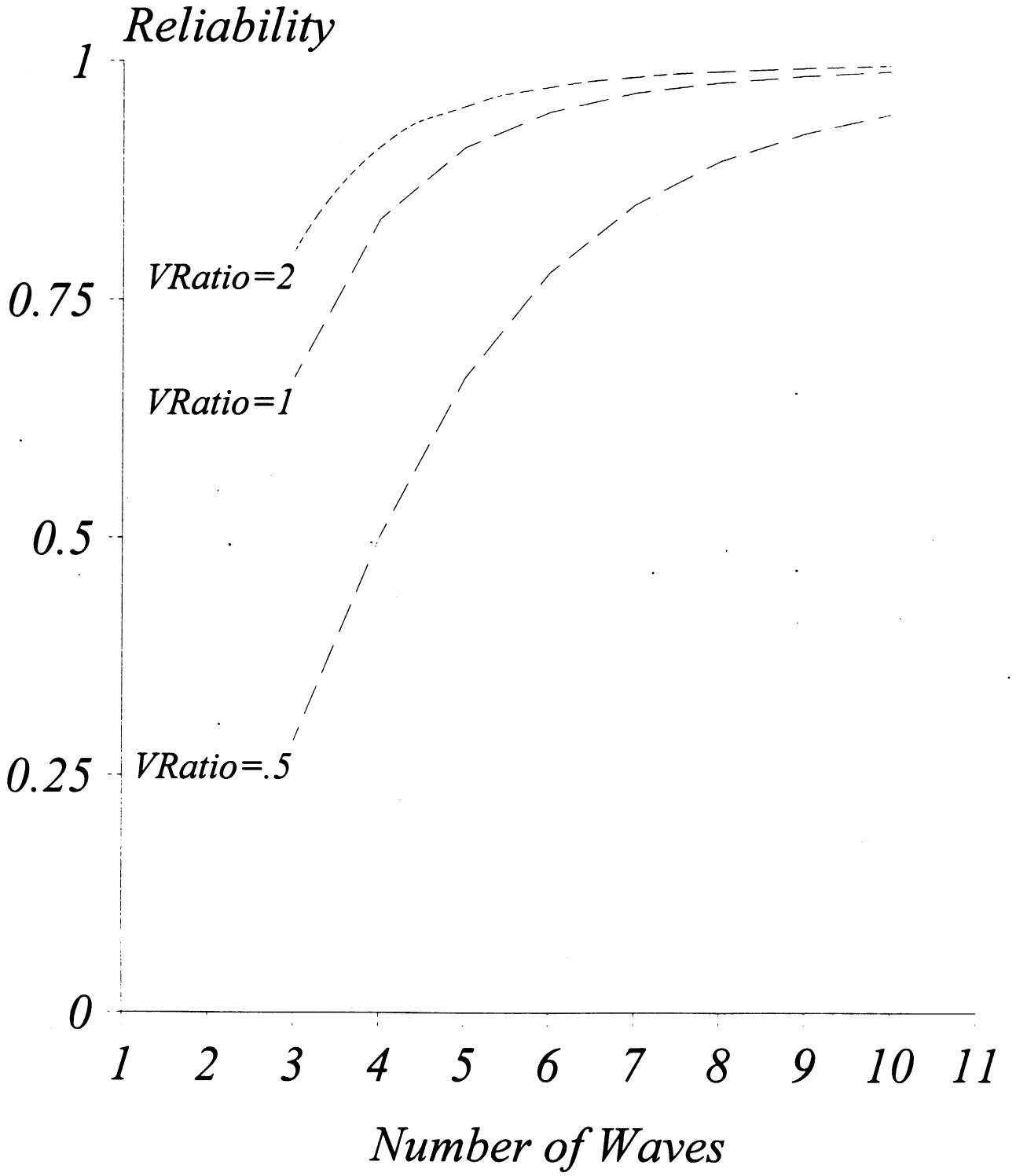
*Standard error of the individual rate of change (in units of residual standard deviation) displayed as a function of the number of waves of data collected*




*Assuming linear individual growth, ordinary least-squares estimation of the rate of change, heteroscedastic residual variance and equally-spaced waves of data.*

Figure 7

*Reliability of the individual rate of change displayed as a function of the number of waves of data collected*




*Assuming linear individual growth, ordinary least-squares estimation of the rate of change, heteroscedastic residual variance and equally-spaced waves of data.*

population.[7] Consequently, in Figure 7, we display the relationship between the reliability of linear change and the number of waves of data at selected values of *VRATIO*, the ratio of the population variance of true linear slope to measurement error variance.

In practice, *VRATIO* can take on widely disparate values. In some domains, measures of a particular construct may be highly fallible and in addition there may be little heterogeneity in true change across individuals in the population. In this case, an individual's observed datapoints will be widely scattered from his or her true growth trajectory, measurement error variance will be high and the population variance of true linear change will be small, leading to a value of *VRATIO* that is much less than unity. The lower curve in Figure 7 displays the relationship between the reliability of linear change and the number of waves of data for the situation in which the population variance of true change is one half the measurement error variance, a situation which is not unusual in practice (see Williamson, Epanchin & Appelbaum, 1992). Notice that, in this situation, when only three waves of data have been collected, the reliability of linear change is quite low (.28).

In other domains, measurement of the construct may be quite precise causing each individual's observed datapoints to fall close to his or her underlying true growth trajectory, and, in addition, population heterogeneity in true change may be large. In this case, *VRATIO* will be larger than unity. The upper curve in Figure 7 displays the relationship between the reliability of linear change and the number of waves of data for the situation in which the population variance of true change is twice the measurement error variance. Notice that now, when only three waves of data have been collected, the

---

⑦ Under the assumptions listed in text, the population reliability of the ordinary least-squares estimate of the linear rate of change is given by:

$$\rho(\hat{\pi}) = \frac{(\sigma_\pi^2/\sigma_\epsilon^2)}{(\sigma_\pi^2/\sigma_\epsilon^2) + \frac{12}{T(T^2-1)}}$$

where $\sigma_\pi^2$ is the population variance of the true linear growth rate, $\sigma_\epsilon^2$ is the population measurement error variance, and T is the number of waves of data collected (see Willett, 1988).

reliability of linear change is very reasonable (.80).

The center curve in Figure 7 is plotted for a value of *VRATIO* that corresponds approximately to the longitudinal measurement of teacher satisfaction, measures of which were constructed by matching four items from the 1987-88 SASS with corresponding items on the companion 1988-89 TFS.[8] In this case, with three waves of data collected in the design, the reliability of change is moderate (.67).

Inspection of the curves in Figure 7 leads to three plausible conclusions relevant for the design of a future longitudinal study of the teacher career:

1. ***The reliability with which linear change can be measured is higher for designs in which more waves of data are collected.*** When three waves of data are collected in the context of our example on teacher satisfaction, for instance, the reliability of change increases to .91 by the addition of two extra waves of data (all else being equal).
2. ***The impact of adding more waves of data collection to the design is greater for designs that initially contain fewer waves.*** In the case of teacher satisfaction, for instance, the reliability of change increases by 38% (i.e., from .66 to .91) when the number of waves collected is increased from three to five but by only a further 5% (from .91 to .97) when the number of waves is increased from five to seven.
3. ***The impact of an increase in the number of waves of data is disproportionately felt in designs in which VRATIO is small.*** In designs in which the measurement of a construct is highly fallible (i.e., in which measurement error variance is high) and population heterogeneity in change is small, the addition of a few extra waves of data will lead to dramatic improvements in reliability. For instance, when the number of waves of data collected is increased from three to five, the reliability of change increases by an amazing 133% (from .28 to .67) when VRATIO is

---

® We are not confident, however, in the quality of our estimates since the specific wording of the teacher satisfaction items and the definitions of their rating scales differed considerably between the SASS and the corresponding TFS.

.2, and by a paltry 19% when VRATIO is 2.

## 5.2 For how long should teachers be observed?

We have argued above, in the introduction to this section, that there are two types of research question about time, in which time can be either conceived as a "predictor" or as an "outcome" (Willett & Singer, 1989). Investigators who ask how much time must pass before a teacher experiences a critical event are implicitly treating time as the outcome and selected characteristics of the teacher's background, environment and treatment as predictors. Such questions are typically answered by the methods of *survival analysis* (Singer & Willett, 1991, 1993; Willett & Singer, 1993, 1991). Links between the optimal duration of data collection and the statistical power of survival analyses allow us to make reasonable guesses about how long we must collect data if we are to be able to answer such questions.

Conceptually, the answer is simple--the duration of a longitudinal study must be long enough to ensure that substantively important changes are observed during the period of data collection. However, teachers are not students--the changes that teachers exhibit in their professional lives are often subtle and slowly paced. Attrition rates from the profession are low and, in addition, teachers move between schools and school districts. Yes, there are pockets of great turnover and change and increasing percentages of teachers are nearing retirement, but as a whole, the nation's teaching force is relatively stable in comparison to years past (Darling-Hammond & Sclan, in press). The study of change and transition in such a relatively stable environment requires the collection of lengthy longitudinal data over a longer period of time. In addition, it is only when such long-duration empirical records are in our possession that we will be able to address important policy questions about the *return* of former teachers to teaching. The reason is simple: for a longitudinal study to describe effectively the return of former teachers to the profession, then data must be collected over a sufficiently long time-period for an analyzable subsample of teachers to return. We reconsider all of these issues below.

Of course, if there were no need for immediate results and money was no object, NCES might

reasonably collect longitudinal data on teachers throughout the *entire* duration of their teaching careers. In theory at least, NCES could follow a sample of beginning teachers from their first year on the job through eventual retirement from the profession. Such a "womb to tomb" research design would be ideal for studying change over the entirety of the career. It would allow us to learn, for example, whether career patterns and levels of commitment differ for teachers who were certified via an alternative route in comparison to other teachers who were certified via a traditional mechanism. Or whether teachers we might have been able to identify, on the basis of early signs and signals, were likely to leave teaching, never to return.

Given a sufficiently frequent set of temporal observations, womb-to-tomb data collection has another important appeal--it enables the researcher to observe *change and transition as they occur.* Thus, if NCES were to collect data frequently throughout the entire teaching career, then crucial times of change or transition would not be missed and a complete record of time-varying longitudinal information on the teaching career would be available to future investigators. This record would then provide a basis for subsequent individual growth modeling and measurement of change and would also be the source of time-varying covariate information that could be used to predict transition occurrrence and spacing (Willett & Singer, 1993).

None of the national data on teachers' careers currently available to researchers provides sufficient longitudinal information about when and why changes occur, and about time-varying covariates, in a representative and up-to-the-minute sample. Clearly, a new longitudinal study of the teaching career is required. However, it seems unlikely that the "womb to tomb" design--the optimal analytic option--will appeal to NCES for reasons of finance and practicality. So, in order to warrant such a study, we must ask ourselves: What is the *minimum duration* for which we might implement such a study in order to encounter enough professional transitions and sufficiently large amounts of change in the major domains?

One way to answer such questions is to examine the rate at which important professional

transitions currently occur in the teaching career and base estimates of the potential duration of the future study on that information. Here again we appear to lack the very data that we need to help make the design decision--we do not possess the current, nationally generalizable, longitudinal information on the teaching career that we need to design and implement the new study. To circumvent this impasse, we have constructed a variety of informative descriptive statistics from the most current national information available to us, the 1987-88 SASS and its companion TFS. Because these surveys provide only two waves of data collected one year apart, the descriptive summaries cannot really be considered to be truly longitudinal. However, they do provide a basis for the ballpark decisions that we must make.

In Figure 8, we present one such summary--pseudo-longitudinal estimates of discrete hazard functions describing the risk associated with a teacher's: (a) leaving teaching, and (b) staying in teaching but moving from one school to another, in each year of a teachers' first, second and third (or higher) spells in teaching.[9] Readers must be cautious in their interpretations of these hazard functions, however, because these profiles are not true discrete-time hazard functions. They do not describe the declining temporal risks that a randomly-selected teacher has experienced as he or she moves from one year of the career to the next. Rather, they display the risk of leaving teaching (and of moving schools) for teachers at all levels of experience in the current spell. Specifically, we constructed the risk profiles in piecemeal fashion from pairs of two-wave datapoints available in both the 1987-88 SASS and 1988-89 TFS on teachers at all levels of experience. For instance, by comparing the responses of teachers in the first year of service of their first spell on both the SASS and TFS, we were able to estimate the probability that a first-spell/first-year teacher would leave teaching in 1988. Similarly, by examining the linked SASS/TFS responses of first-spell/second-year teachers, we estimated the risk associated with the second year of first-spell service in 1988, and so forth. By plotting these estimated probabilities against years of

---

⑨ The discrete-time hazard associated with the occurrence of a critical event is defined as the probability that a randomly-selected member of the population will experience the event in any discrete time-period, given that he or she has not already experienced it (Singer & Willett, 1993).

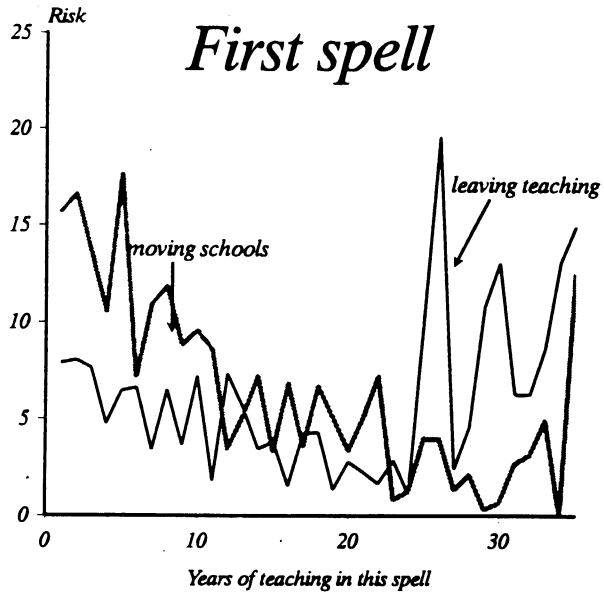
teaching in each spell, we have produced a pseudo-longitudinal display of the risk of leaving teaching in every year of experience in each spell. Similar piecemeal computations were carried out for the subsamples of teachers in the second and third spells in the profession, and for the equally-important event of moving from one school to another. With these provisos in mind, we will base subsequent discussion on these important summaries.

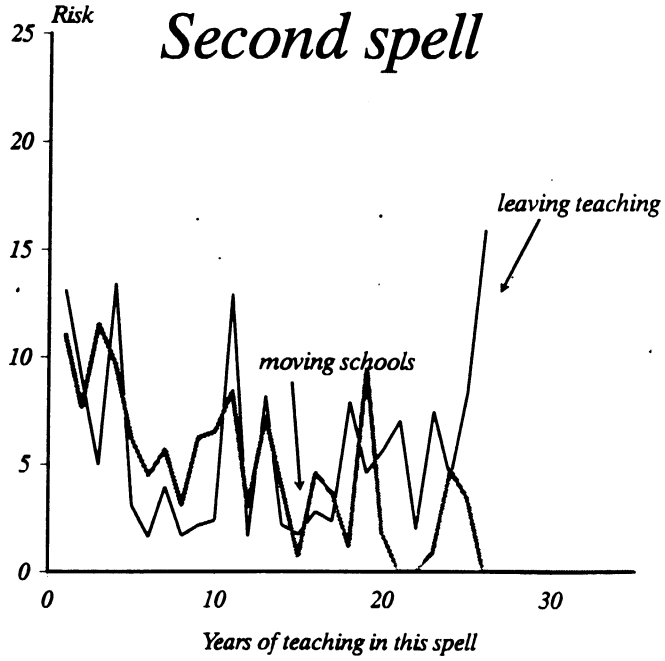
Despite their obvious drawbacks, the pseudo-hazard profiles in Figure 8 provide us with evidence about the rate at which teachers make transitions during the various stages of their professional careers. We will consider teachers in each of the three spells separately. For teachers in the first-spell, inspection of the first panel in Figure 8 suggests that the career can be divided into three epochs of roughly-equal duration: (a) the early years (0-12 years), (b) the middle years (12-24 years), and (c) the later years (25-36 years). Risks in the first epoch tend to be elevated initially but decline with the passing years--a phenomenon found in most other studies of the teaching career (see, for instance, Murnane, et al., 1991). One critical observation that we might make, however, is that, during the first epoch, *teachers experience a much greater risk of moving from one school to another than they do of leaving the teaching profession entirely.* In each of the first three years of first epoch, for instance, approximately 15% of the teachers will move to another school, whereas only 8% will leave teaching entirely. As years pass in the first epoch, risks of "moving versus leaving" converge so that, by year twelve, about equal proportions of teachers will both leave and move. During the middle years of the first spell, this risk-equality is approximately preserved until the risks of leaving teaching dramatically increase in the third epoch (years 25 through 36) with the onset of retirement.

In the second and third spells, the U-shaped pseudo-hazard profile of the first spell is again found but epochs within the career are less distinct and, since sample sizes are necessarily smaller, evidence of sampling variation is more apparent. Again, however, risks appear to be somewhat elevated early in the spell. About 10% of the teachers both leave and move in each of the years 1 through 5 of the second spell, whereas less than 5% leave and move in each of the years 5 through 10. Again, as in the first spell,

Figure 8

# *Estimated risk of moving schools or leaving teaching*



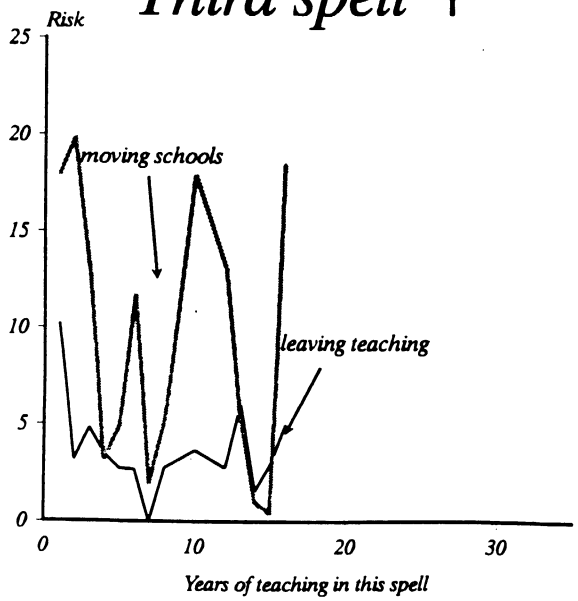
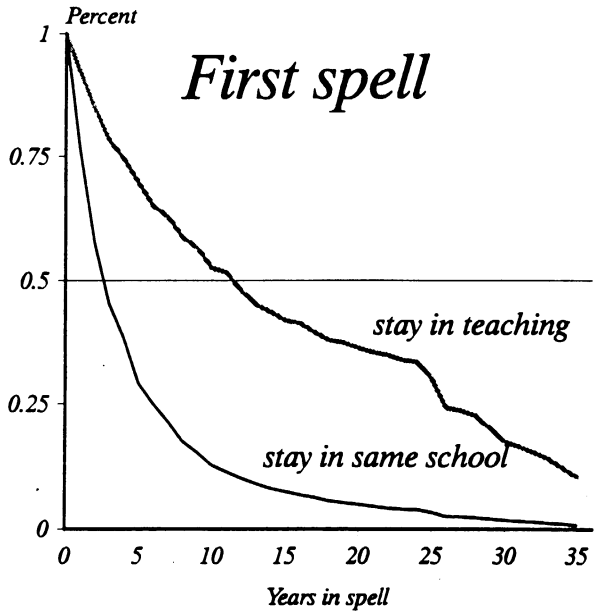
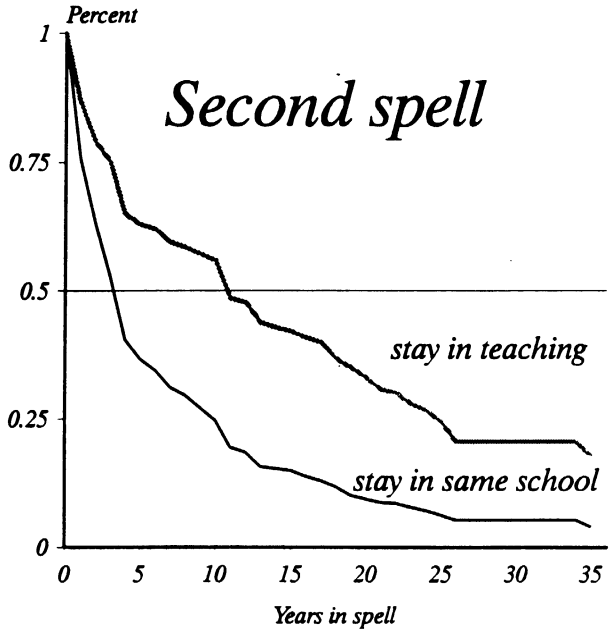
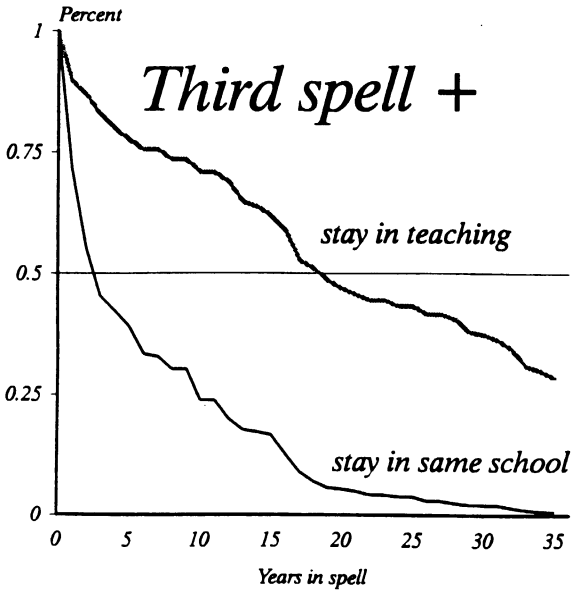



Figure 9

# *Estimated survival functions for staying in teaching and staying in the same school*



we observe convergence of the risks associated with both moving and leaving in the middle years of the second spell and a subsequent increase in the risk of leaving the profession with the onset of retirement. It is difficult to reach firm conclusions about the risks associated with the third spell due to the small sample sizes involved and the resulting imprecision in the obtained risk estimates.

Despite their imperfections, the pseudo-hazard functions in Figure 8 provide important guidance for research design. First, we have noted that most of the "action" occurs in the early years of each spell, where risks of both leaving and moving are elevated. This suggests that, if we wish to design a longitudinal study so that transition and change in the teaching career are observed, we must focus particularly on the early years of the career for teachers in both their first and second spells, if not also their third (and higher). This corroborates our earlier design decision to use first-year teachers, or beginning teachers, as the target population for the proposed longitudinal study. Second, given the relative prevalence of "moving" over "leaving" in the early years of the career (particularly in the first spell), we must design any new prospective study of the teaching career to ensure that the former transition is awarded at least as much attention as the latter. This implies that any new study must track teachers as they move from school to school, from school district to school district, and from state to state. These types of transitions have been largely ignored in current research on the teaching career, usually because data are not available on teacher movements between states. This then represents an important niche that NCES could easily fill with its new study.

But for how long should we plan to observe these teachers? We can obtain estimates of prospective study duration by examining a second set of pseudo-longitudinal summaries obtained from the SASS and the TFS--the discrete-time survival functions associated with each of the hazard profiles in Figure 8. We provide these pseudo-survival functions in Figure 9. For reasons that we have already discussed, while they can be regarded as cumulating the risks associated with moving and leaving that were displayed in Figure 8 as is typical of a survivor function (Willett & Singer, 1993), caution must be

exercised in interpreting these new summaries as they are not true survivor functions.[10]

The three panels of pseudo-survivor functions in Figure 9 magnify the effects that we have already noted in Figure 6 and support our earlier statements about the relative prevalence of "moving school" over "leaving teaching." In each panel, we note that the pseudo-survivor function describing the probability the a randomly-selected teacher will stay in teaching is always consistently higher than the corresponding function describing the probability that the teacher will stay in the same school. In other words, in every year of the career, teachers are more likely to survive in *any* school than in *any particular* school. Inspection of the pseudo-survival functions also suggests that most of the dramatic changes in survivorship are likely to take place within the first ten to twelve years on the job, the functions plummeting rapidly in these early years.

We can compute how many years of experience must pass before the average teacher either leaves teaching or moves to a different school in his or her first, second and third spells by estimating the median lifetimes associated with each of the pseudo-survivor functions in Figure 7. We simply extend a horizontal line on each plot at a survival probability of 50% and impute the number of years of experience that corresponds to these probabilities in each spell. This provides an estimate of how many years of experience must pass before half of the sample has experienced the event of interest. We list these estimated median lifetimes in Table 1.

Notice that, regardless of spell, there are considerable differences between the median lifetimes associated with a teacher switching schools and leaving teaching, regardless of spell. On average, and over all spells, teachers tend to switch schools quite readily--the average new teacher switching to a new school in just less than 3 years. It takes somewhat longer for the average teacher to decide to quit the profession entirely, the median lifetimes here ranging from about eleven years in the first and second

---

[10] In any given time-period, the discrete-time survival probability associated with the occurrence of a critical event is defined as the probability that a randomly-selected member of the population will experience the event of interest after the time-period in question.

Table 1. Estimated median lifetimes in years for teachers in their first, second and third spells, for the events of: (a) switching school, and (b) leaving teaching. All estimates are imputed from the pseudo-survivor functions in Figure 3.

| Spell number | Average number of years in spell before... | |
|---|---|---|
| | Switching schools | Leaving teaching |
| 1 | 2.6 | 11.4 |
| 2 | 3.2 | 10.8 |
| 3 | 2.5 | 18.4 |

*This page intentionally left blank.*

spells to slightly more than 18 years in the third spell.

Knowledge of the median lifetimes anticipated among teachers in the different spells permits us to use the method of power analysis to suggest an appropriate duration for the proposed study of teachers' careers. Although no single power analysis can cover all possible design configurations, by making reasonable simplifying assumptions we can obtain ballpark estimates of duration that provide concrete and reasonable guidance for the proposed study. For instance, in an investigation of the study duration and minimum sample sizes required to detect effect sizes of 1.5 or 2 (these effect sizes being typical in the study of the teacher career--see Murnane, et al., 1991), Singer and Willett (1991, Table 1, p. 277) show that the minimum sample size required to achieve a statistical power of .80 varies considerably as a function of proposed study duration.[11] They stress that, when follow-up extends to the median lifetime, then the same power can be achieved with almost half as many individuals as would be required if follow-up extended to only half of the median lifetime. If follow-up is extended to twice the median lifetime then the same power can be achieved with only a third as many individuals. The message for research design is clear--much statistical power can be gained by following study participants for longer periods of time.

As a result of their comparisons, Singer and Willett (1991) conclude that study participants should be followed for at least as long as the median lifetime, and preferably longer. Thus, based on the anticipated median lifetimes listed in Table 1 above, we recommend that, in the new study that is being planned, NCES should observe entering cohorts of teachers in their first, second and third spells for at least 12 years. Not only will such a design be likely to provide adequate statistical power at reasonable sample sizes, but it will also ensure that approximately 50% of teachers will leave teaching during the period of observation, thereby contributing to the healthy pool of potential returnees available for study.

A further implication of the median lifetimes listed in Table 1 above, is that sampled teachers

---

11. In the computations underlying their table, Singer and Willett (1992) assume a two-group comparison at the .05 level (two-tailed) with a uniform population hazard function.

must be followed as they move from school to school. On average, from information contained in the SASS and TFS, we can infer that teachers tend to switch schools approximately every three years. Thus, the selection of a 12-year observation window will not only ensure that moderate numbers of teachers will quit their jobs during the period of observation, but also that very large numbers of teachers will switch schools during the same period. This makes it particularly important to follow these teachers from school to school, collecting time-varying information on their situation, treatment and behavior.

### 5.3 What is our recommendation?

We have argued that there can be no clearcut answer to the twin questions of "how long?" and "how often?" other than "it depends." However, based on the arguments in this section, we are prepared to make what we believe to be reasonable suggestions--NCES should plan on **at least twelve years of observation and six waves of data collection equally-spaced throughout the period of observation**. Based on our arguments about the links between statistical power and study duration, an observation window of twelve years would almost certainly ensure that about half of the teachers in the sample will experience the critical event of "leaving teaching" during the period of interest, and many more would make the transition from one school to another. Based on our arguments about individual growth modeling, obtaining six bi-annual observations on each teacher would ensure that growth of limited curvilinearity could be detected and that individual change could be measured at reasonable levels of precision and reliability.

## 6.0 Issues of implementation:

## Implications of other methodological concerns for the conduct of the longitudinal study

Many of the methodological issues that we have raised have important ramifications for implementation. Although implementation issues are often considered *after* core design decisions have been made, it is clear to us that the success of this longitudinal study rests squarely on the resolution of several key issues of implementation. For this reason, we now take the opportunity to address briefly seven issues that we believe NCES needs to consider *before* fielding their data collection effort. These inclued: (a) replicating the study in multiple base years, (b) expanding the modes of data collection, (c) carefully reconstructing retrospective event histories, (d) equating measurements across occasions, (e) tracking teachers in and out of school, (f) embedding substudies of natural experiments in the larger study, and (g) piloting design features before embarking on the full longitudinal investigation.

### 6.1 Replicating the study in multiple base years

In section 3, we strongly suggested that NCES consider conducting the study in *several* base years. The primary reason for doing so was to help researchers unravel the age-period-cohort problem. By selecting a sample of first year teachers who began teaching in each of several entry cohorts, researchers would eventually be able to determine whether the behavior of the first year teachers in the initial sample was a consistent feature of first year teachers, or was possibly a feature of first year teachers hired *in that particular entry cohort.*

NCES already recognizes the virtues of including multiple cohorts in several of their data collection programs. The SASS and TFS are fielded on a periodic basis, as are the multiple studies of high school students. As we suggested under design principle #5, there are strong incentives for replicating this study at several time points after the initial implementation. But we hasten to add that

the sample size for each initial target population must be sufficiently large to permit statistical analysis with ample power.

Because this decision can be thought of as secondary to the decision to mount a longitudinal study in the first place, we do not spend much time here discussing the distances between initial cohorts. It is virtually impossible to predict what first year teachers will look like three years from now, let along five or six years from now, when a second initial cohort might begin. Rather, we simply remind NCES that there are great gains to be made by initiating the longitudinal study in several base years, and that if they decide to go forth with this planned longitudinal study, they consider this possibility in their design.

### 6.2 Modes of data collection

As presently implemented, the SASS and TFS use mail surveys conducted by the Census Bureau for NCES. Questionnaires, sent to teachers, principals, and district personnel, are returned to the Census Bureau for data processing. Mail surveys, amongst the least expensive data collection modes available, have been adequate because the types of information that NCES has heretofore chosen to collect have been answers to either factual questions (e.g., in what year did you start teaching in this school?) or a small number of attitudinal items (e.g., how satisfied are you with your salary?). Reinforcing these measurement decisions has been the result of reinterview studies, in which NCES has found that mail questionnaires produce fewer inconsistent responses than do telephone questionnaires and that factual items, not surprisingly, are more likely to be answered consistently in contrast to attitudinal items.

The data needs outlined in the first section of this paper suggest that NCES will need to augment these mail surveys with other modes of data collection. We believe that it will not be possible to gather information on some of the constructs of greatest interest--on teachers' worklives, teacher quality, teacher work contexts, and teachers' career paths--solely using data that can be collected as part of a mail survey. Although we do not want to delve into the substantive specifics of which items should be asked using which data collection approach, it is clear to us that in planning for this longitudinal study, NCES

must consider the merits of using telephone interviews, diaries, in-class observations, administrative records, and links with other NCES surveys. Different methods may be required for the measurement of different underlying constructs.

To illustrate these ideas, we briefly consider the advantages of adding a series of measures designed to assess time use--how teachers spend their time in school and how they spend out-of-school time preparing for school. Although the NEA collects such data in their periodic *Status of the American Public School Teacher* surveys, NCES has shied away from collecting detailed data on this topic. Issues about time use and time allocation clearly deserve investigation, as this is an aspect of workplace conditions that is likely to vary over time. Although it might be possible to have teachers answer questions about these topics using a grid provided in a mail questionnaire or by responding to a series of guided questions in a telephone survey, a diary methodology might provide higher quality data (Silberstein & Scott, 1991). Diaries, a common data collection tool in expenditure surveys, do not play as prominent a role in NCES studies. Time use data collected by diaries could serve as both outcomes and predictors. Researchers could use these data to address questions such as whether time allocation changes as teachers gain experience or whether teachers who are given few preparation periods are more likely to leave.

Some of the alternative data collection strategies that we mention above would provide data so useful that NCES might choose to use them for *all* respondents. NCES might consider, for example, conducting a brief telephone interview every other year in between the major biannual assessments to determine whether the teacher is still teaching, and if so, where. But recognizing that most alternative forms of data collection require increased resouces--both time and money--NCES might consider embedding some of these more intensive strategies in a *smaller subsample* of the larger study. Few NCES data collection efforts include embedded substudies, yet it will not be possible to include measurement approaches such as in-class observations for an entire national probability sample. There is a great deal of merit in treating the decreased generalizability that would come from an embedded

substudy as a reasonable price to pay for the increased focus that could result from such measurements. Better to have some longitudinal information that actually measures the constructs which we would like to measure on a subsample of teachers than less detailed information on the entire group. The types of questions currently asked on the SASS and TFS (at all respondent levels) simply do not delve into the underlying constructs deeply enough to get at the types of information researchers require for adequately investigating the teaching career.

Freed from the constraint of having to administer the full complement of measures to *all* sampled teachers, NCES could augment the basic data collection scheme with a variety of innovative data collection approaches. The myriad possibilities include: (a) a more frequent data collection schedule for subgroups of teachers; (b) in-class observations in selected geographic areas; (c) distributing "work diaries" in which teachers would record how they spent their school time during a two-week period, perhaps with the distribution plan for the diaries constructed so as to cover an entire school year; (d) conducting telephone interviews with principals and teachers several times during the school year; and (e) linking SASS data collection with NAEP data collection or state-level administrative records on salaries, benefits, and work histories. The epidemiological literature provides much support for this last idea. In their on-going data collection efforts, for example, NCHS has linked data from the Health and Nutrition Examination Surveys to death records and has linked data from the National Medcial Care Use and Expenditure Surveys to Medicaid records.

### 6.3 Retrospective reconstruction of event histories

The retrospective reconstruction of event histories is one measurement topic we would like to single out for detailed focus. We do so because we believe that the present approaches NCES uses are inadequate for getting at the types of information needed, and because retrospective reconstruction of event histories will continue to play a major role in this planned *prospective* data collection effort. To be sure, focusing the longitudinal study on teachers in the first year of their current spell will alleviate some

of the measurement burden involved in the retrospective construction of event histories. Approximately two thirds of the sampled teachers will be in their first spell and for them, the problem of retrospective data collection should be relatively minimal.

But even for teachers in their first year of their first spell, there will be a need to collect salient employment history information. And for the remaining one-third of teachers who will have taught at some point in the past, there is a pressing need to collect data that will allow researchers to retrospectively reconstruct their career paths in full. Only with this information will researchers be able to understand how returning teachers behave over time.

Why do we focus so closely on the retrospective reconstruction of event histories? One reason is that we have found that the current questions included in the SASS are inadequate for reconstructing career histories in sufficient detail to even help with planning for this future longitudinal study. In some questions, teachers are asked to supply the *year* in which a particular event occured (when did you first begin teaching and when did you first begin teaching in *this* school) while for other questions, they are asked to supply to *total number of years* that they have been in a particular state (for how many years have you been teaching full time in public schools). And when it comes to breaks in service, teachers are asked only to indicate the total number of breaks, not their distribution. Although we cannot provide evidence about telescoping and memory lapses, we have already seen ample evidence of rounding errors. A quick glance at the peaks on the distributions of years in spell in Figure 4 shows, for example, that teachers are more likely to supply answers that end in either 5 or 0. More precise information will be needed for the future longitudinal study.

A second reason that we focus on improving the quality of retrospective data collection is because we believe that in some ways, almost *all* of the information collected in this planned *prospective* study will actually be *retrospective*. When teachers are asked such questions as whether they have taken an in-service course during this school year, whether they have participated in school based decision-making, and whether violence is a problem in their school, teachers must reflect on their past experiences

and attempt to reply as reliably and validly as possible. Yet few of the questions about teachers' experience currently include a time frame for reference. Without an explicit reference, we do not know whether teachers are describing the school as they see it "today," "during this school year," or "in comparison to what it once was."

What is the value of retrospectively reconstructing event histories? Many would argue that it is time-consuming and costly to collect such data, and when the data are in, they are of questionable reliability and validity. In an article describing the circumstances under which researchers may appropriately analyze data on "backwards recurrence times," the time elapsed since a last event such as leaving teaching or moving schools, Allison (1985) writes:

> It is reasonable to ask whether such data [time since last event] should continue to be collected or whether more effort should be expended in getting complete event histories. Certainly more complete data are always desireable when it is practical to collect them. On the other hand, eliciting accurate data on long histories of events can be a difficult and costly process. (p. 322).

Although we agree that retrospective reconstruction is expensive, for this planned longitudinal study, we believe that it is also imperative. Of particular concern is the prior teaching experience of returning teachers and the prior employment experience of beginning teachers. If NCES cannot devote sufficient resources to reconstructing these previous employment histories, we will have no way of placing future work behavior in a personal context. We would go so far as to say that if NCES cannot measure the previous teaching experience of returning teachers, they would be better off setting these teachers aside from the target population and focusing exclusively on first year teachers in their first spell.

Can reliable and valid retrospective data be gathered? Sociologists and psychologists have made great strides towards understanding why it is difficult to recall past events and towards developing improved methods for reconstructing event histories. For example, it is now clear that improved responses can be gotten if questions about *when* an event occurred are linked to questions about *where* and *why* the event occurred (Bradburn, Rips, & Shevell, 1987; Featherman, 1980; Friedman, 1993; Wagenaar, 1986). Although NCES' primary interest may be in event *timing*, better information about

timing can be had if the questions asked follow a more narrative format with cues designed to aid recall. NCES should consider working with cognitive survey measurement experts on designing a protocol that will elicit teachers' previous teaching history, beginning with student teaching experiences and extending to their present job. For teachers in their first spell, the protocol will be relatively brief; for teachers in later spells, the interview should be flexible enough to collect as much data as is necessary to get a complete employment history. Although NCES will probably not be able to gather retrospective data on time-varying covariates (Gutek, 1978; Bradburn et al., 1987), improved event history data could help researchers place returning teachers' future career behavior in the relevant broader context.

The need to improve the data collection strategies for eliciting a description of previous work experience can be seen in the current questions on the SASS. Early in the instrument, teachers are presented with a confusing series of questions organized in varying levels of recency (from first job to current job to previous job) and elicited in varying metrics (some questions ask for calendar years; others ask for elapsed time). Among the many issues that will arise in the redesign of this protocal are whether: the data should be collected by mail questionnaire or telephone interview; the instrument should move forward in time (from first job to most recent job) or backwards in time (from most recent job to first job); the cues should be a characteristic of the school or of the teacher's job assignment; and time should be measured in calendar years or school years. In reviewing questionnaires that might be used to help design this revised inventory, we recommend that NCES staff examing two interview protocols that we have found particularly helpful: one for smoking history developed by Means, Swan, Jobe and Esposito (1991) and the life history calendar developed by Freedman et al. (1988) to assess residence, marital, and employment history.

### 6.4 Ensuring measurement equatability

In any longitudinal study, the issue of measurement equatability is paramount. Whether we wish to model individual change over time in a particular construct or to use that construct as a time-varying

covariate in an event history analysis, we must ensure that all measurements of the construct are valid and in an identical metric, regardless of the occasion of measurement. During data analysis, it is impossible to model change in a construct or to make sensible interpretations of the effects of a time-varying covariate if we have measured apples on one occasion and oranges on the next.

The requirement for measurement equatability seems so obvious and straightforward that it is difficult to understand how violations of equatability can occur in well-planned longitudinal studies. But they do occur, and frequently! Even in the 1987-88 SASS and the companion 1988-89 TFS, all items measuring teacher satisfaction were modified considerably between the two admininstrations of the survey leading to construct-wide non-equatability of all items. For instance, with regards to teachers' satisfaction with the level of material resources available to them, item 29(h) of the 1987-88 SASS asked teachers:

> Do you agree with the following statement?
>
> > Necessary materials (e.g., textbooks, supplies, copy machine) are available as needed by the staff.
> >
> > 1 = Strongly agree
> > 2 = Somewhat agree
> > 3 = Somewhat disagree
> > 4 = Strongly disagree.

Then, the following year on the 1988-90 TFS, in a so-called "equivalent" item (#27.12) teachers were asked:

> How satisfied are you with EACH of the following aspects of teaching? Are you (a) Very satisfied, (b) Somewhat satisfied, (c) Somewhat dissatisfied, or (d) Very dissatisfied with --
>
> > Availability of resources and materials/equipment for your classroom.

Notice that, in addition to phrasing the item stem differently between the two administrations, the

response categories were also modified. In the SASS, respondents were asked to indicate their level of *agreement* with a statement about whether specific materials were available when needed by staff. In the TFS, on the other hand, respondents were asked to rate their *satisfaction* with a more general specification of resources, materials and equipment. Because of these changes, even though responses on both occasions were coded on a four-point scale, the data analyst cannot assume that scores obtained on the two occasions are equatable. It seems so wasteful to have intended to measure the same construct--aspects of teacher satisfaction--on each of two occasions and then be unable to investigate change over time in this construct because of thoughtless and unnecessary changes in phraseology.

To ensure the equatability of measurement over time, it is a good idea to admininster an identical instrument each and every time the construct is measured. Items must not be "updated" on subsequent measurement occasions to eliminate and replace poorly performing items. The item stems and responses must not be adapted or modified, sections of phrasing must not be replaced. Response scales must not be changed from four-point to five-point scales in subsequent editions of the instrument. None of these changes is without cost--even though they may appear harmless to the untrained eye, they will render measurements made with the instrument non-equatable across occasions and prevent sensible subsequent analyses of change and time-varying covariation. The time to modify a measuring instrument is during pilot studies conducted before the main longitudinal study itself.

Of course, there are at least three disadvantages to admininstering an identical instrument repeatedly to the same individuals on many occasions. First, repeated exposure to the same instrument may lead respondents to become familiar with the instrument itself and their memories of prior administrations may cause their responses to be unwarrantedly modified or maintained, or perhaps lead to measurement errors becoming auto-correlated over time. However, under the design that we have recommended for the proposed longitudinal study of teachers' careers--with bi-annual surveys equally-spaced over a period of twelve years--we feel that measurement occasions will be sufficiently spaced out to alleviate these memory-related problems. Second, there is an issue of construct validity--even though

the same instrument may be admininstered on every occasion, it may not necessarily be measuring the same construct on each of them. For instance, in the case of the measurement of arithmetic skills in young children, an instrument intended to measure multiplication skills at age six may have become a measure of rote memorization by age eight. However, we believe that this type of "construct-shifting" is less likely to occur when adults are the subjects of study, and we therefore anticipate that the problem is unlikely to occur in the proposed study of teachers' careers. Third, when respondents are growing over time, it is possible that the administration of an identical measure on several occasions of measurement will lead to the appearance of floor and ceiling effects. In the early years, for instance, when teacher efficacy is low, the measure may not be sensitive enough to register its value and to discriminate among individuals on the basis of their efficacy. Then, in the later years when efficacy has improved, the scale may not be "long" enough to accomodate the scores of individuals who have changed rapidly and whose efficacy is very high, even though the measure may have retained its metric and remained construct valid over time.

We stress that problems of measurement equatability must be anticipated in advance and addressed carefully at the research design stage. They are extremely unlikely to be able to be resolved during data analysis. For this reason, we recommend that NCES address them in advance in a series of well-planned pilot studies, perhaps executed on representative subsamples in conjunction with the SASS and the TFS prior to the main longitudinal study.

### 6.5 Tracking teachers in and out of schools

In section 5 of this paper we concluded that, during the course of their professional lives, teachers were much more likely to switch schools and continue teaching than they were to quit teaching entirely, regardless of spell. We review and summarize the evidence for this conclusion in Table 2 in which we display estimates of the percentages of teachers who will remain either in the same school (columns 2 through 4) or remain in teaching (columns 5 through 7) for 2 years, 6 years and 12 years,

Table 2. Estimated 2-year, 6-year and 12-year survival rates for teachers in their first, second and third spells, for the events of: (a) switching school, and (b) leaving teaching. All estimates are imputed from the pseudo-survivor functions displayed in Figure 3.

| Spell number | Estimated percentage of teachers who have not... | | | | | |
|---|---|---|---|---|---|---|
| | Switched schools after... | | | Left teaching after... | | |
| | 2 years | 6 years | 12 years | 2 years | 6 years | 12 years |
| 1 | 57.5 | 25.1 | 10.2 | 84.7 | 65.1 | 47.8 |
| 2 | 63.2 | 34.5 | 18.6 | 79.0 | 62.0 | 47.8 |
| 3 | 55.3 | 33.5 | 20.0 | 86.9 | 75.6 | 68.8 |

*This page intentionally left blank.*

respectively. Notice that, regardless of spell, between 50% and 60% of all teachers remain in the same school for at least two years but by 6 years into the career less than one third remain. After twelve years, only about one fifth remain in the school in which they began the spell. Compare these percentages to the proportion of the teaching force that quits teaching entirely in the same periods, displayed in columns 5 through 7 of the same table. In the latter case, many more teachers "survive" -- in fact, regardless of spell, by twelve years, about half of the teachers still remain in the profession.

These summary survival rates contain a critical message for the proposed research design and it is that teachers can be expected to move schools, perhaps several times, before they quit teaching. This implies that, if the proposed study is to be successful in revealing the intricacies of the teaching career in any depth, then NCES must ensure that teachers will not be lost to follow-up as they move between schools during their frequent switches of venue, regardless of spell. Indeed, the very fact that teachers are expected to move frequently from school to school, district to district, and state to state, during the study suggests that it may become both intrinsically more difficult to track them over time than is typical in a longitudinal study. For this reason, we advise NCES to take extraordinary pains to remain in touch with the teacher-respondents at all times in order to reduce losses due to attrition. However, we also believe that should attrition occur during the study, the availability of longitudinal data on respondents prior to their point of attrition would allow the occurrence and timing of the event of attrition itself to be modeled by survival analysis and permit the application of selectivity-bias corrections in subsequent analysis.

An additional consequence of the increased frequency with which teachers switch schools is that variables which we might formerly have considered time-invariant may now become more likely to vary over time. In particular, certain contextual aspects of the school and work environment (such as the overall level of available resources or the socioeconomic status of the school's catchment area) that we might have been willing formerly to regard as time-invariant in a less mobile workforce must now be measured repeatedly over time as their values fluctuate when the teacher switches from one school to the

next. The frequent switching of schools on the part of teachers, also underlines that NCES must anticipate large fluctuations over time in the abilities, behaviors and backgrounds of the students whom a teacher serves. In a mobile teacher workforce, student change is no longer simply a function of student maturation, promotion and graduation but is also strongly dependent on the nature of the professional moves that the teacher makes. Thus, it becomes all the more important for the proposed survey to carefully record the time-varying characteristics of each group of students with whom the teacher works. It can only be anticipated that all of these additional temporal fluctuations will place additional demands on the planned bi-annual data-collection.

Finally, although teachers seem to quit teaching much less frequently than they switch schools, we must not forget that about half of all teachers in any spell will abandon the teaching profession for other work during the period of observation. Of these, many will return--usually within one to two years of quitting. In order to address research questions about the career profiles of teachers who have quit and rejoined the profession, we believe that it is critically important for NCES to commit to continuing to survey former teachers while they are out of teaching. This out-of-teaching measurement will be particularly important in the case of important time-varying variables whose values can only be obtained retrospectively for teachers who join the survey in their second and third spells in the base year. Only when these measurements are available will we be able to effectively answer research questions about "Who returns to teaching?"

### 6.6 Embedding substudies of natural experiments in the larger study

The teaching profession, and education in general, is in a period of great transition. Teacher education programs are changing. Requirements for hire are in flux. Teaching standards are being developed by a variety of organizations operating at both state and national levels. States are developing alternative pathways into teaching. Mentoring programs for beginning teachers have been established in many school districts. Add to these reforms the administrative and policy changes taking place in

America's schools--school based decision-making, school choice plans, changing curricula--and even the distant observer sees that these schools are changing and that any longitudinal study must keep abreast of these changes.

One danger of a longitudinal study is that as the cohort of teachers ages, the results can become less and less relevant. This is one reason we have argued for reinitiating the longitudinal study in at least two and preferably three base years (separated by a period of several years). But there are other dangers as well. If the data collected do not register these changing policies, the amount of unexplainable variation will escalate. Without knowing that a district has implemented a first-year teacher mentoring program, for example, what might explain a finding of great colleagiality and interaction reported among the teachers in a particular school district?

Rather than treat the existence of many divergent policies as a nuisance, NCES could be proactive and regard these varying policy initiatives as an opportunity to collect longitudinal data on the teachers involved in these settings. True, these data could not be used to evaluate the efficacy of these alternative programs for random assignment of teachers to programs seems unlikely. But given that the programs are going forth regardless of NCES' data collection schedule, it would be possible to collect data on teachers participating in such programs and describe their natural course of development.

We do not recommend that NCES compromise the design of this future longitudinal study in adding these embedded components. The core set of respondents should be selected without attention to participation in such innovative programs. But there are other options. In particular, we recommend that NCES consider: (1) collecting data from the teachers, principals, and districts on the implementation of these alternative programs; and (2) augmenting the sample by adding samples drawn from particular districts and states in which new innovative programs were being fielded. If NCES believes that the second option is beyond the scope of their responsibilities, they might be able to work with state and school district personnel to mount a parallel substudy that would enable the collection of the requisite data.

The key point here is that such innovations are going on all over the country whether NCES collects data on them or not. NCES could adopt the stance that these innovations are simply part of the natural variation in policies that exists in American education and that it is not necessary to single out any particular innovation for intensive study. Alternatively, NCES could view this natural variation as an opportunity, and we believe that there is the potential, at least, for sufficient information to be gained that warrants serious consideration of this idea.

### 6.7 Pilot studies

We conclude this section with a brief plea for what we believe is probably the most important implementation issue for NCES to consider: the need for pilot studies in advance of fielding this planned longitudinal study. In writing this report, it became clear to us that answers many of the important design questions require information that is not currently available to either NCES or the research community at large. To their credit, NCES currently conducts pilot studies on a host of implementation issues. For example, studies have compared teacher reports of degree completion with transcript reports. Other studies have compared responses to telephone interview with responses to mail questionnaires. Still other studies have performed reinterview checks, determining whether the questions used display sufficient evidence of test-retest reliability.

Yet many questions remain. How can data on teacher quality be gathered? How variable are teachers' reports of their attitudes towards their jobs? Are attitudes more variable than behavioral indicators of how teachers spend their work hours, or are they less variable? Can a teaching history interview be developed and pretested to determine whether the responses given are reliable and valid? To field a longitudinal study without investing the necessary resources in pilot testing would risk the ultimate waste of large sums of money. Thus, we believe that if NCES wants to take the next step towards developing the design for a longitudinal study of teachers, a series of pilot studies needs to be planned. In this regard, we cede the detailed consideration of the particular pilot studies that need to be conducted to NCES staff and substantive experts.

## 7.0 Conclusion

In this report, we have outlined our view of the pressing methodological issues inherent in the design of a longitudinal study of teachers' careers. We have tried to be broad--describing general goals and design principles--and specific--providing our best judgments about design and implementation issues ranging from the specificity of a target population and the length of data collection to the measurement tools that would be required for a future longitudinal study.

We recommend that NCES focus its energies on beginning teachers--or perhaps more specifically, on first year teachers--who may be in any particular spell of the teaching career. These teachers should be followed for approximately 10 to 12 years with a major data collection effort at least every two years, providing between 5 and 6 waves of longitudinal data. NCES should investigate the possibilility of collecting some types of data on a more frequent schedule, and they should also consider broadening out the types of data collected and the measurement tools used. A longitudinal study of teachers' careers adhering to these guidelines would provide an invaluable, and currently non-existant, research base for the education community to learn more about those individuals so central to the education enterprise.

We believe that if the proposed longitudinal study adheres to these principles, it will represent a sustantial leap forward from our current state of knowledge. Unlike studies based solely on administrative records, for example, this study would collect data from the teachers themselves. Unlike the analyses of national probability subsamples of teachers included in other large national surveys (such as the NLS-72 and the NLSY conducted by the Center for Human Services Research at Ohio State), this longitudinal study would gather data on a range of informants so that we may better understand teachers worklives in context. Unlike the current SASS and TFS, this study would be truly longitudinal, not relying solely on retrospective reports of previous career decisions and limited one-year prospective reports.

## 8.0 References

Akerlof, G. A., & Main, B. G. M. (1979). Pitfalls in markov modeling of labor market stocks and flows. *The Journal of Human Resources, 16,* 141-151.

Allison, P. D. (1985). Survival analysis of backward recurrence times. *Journal of the American Statistical Association, 80,* 315-322.

Billingsley, B. S. (in press). Teacher retention and attrition in special and general education: A critical review of the literature. *The Journal of Special Education.*

Bobbitt, S. A., Faupel, E., & Burns, S. (1991). Characteristics of stayers, movers, and leavers: Results from the teacher followup survey, 1988-89. Washington, DC: National Center for Education Statistics.

Bradburn, N. M., Rips, L. J., & Shevell, S. K. (1987). Answering autobiographical questions: The impact of memory and inference on surveys. *Science, 236* 157-161.

Brookmeyer, R., Day, N., & Pompe-Kirn, V. (1985). Assessing the impact of additional follow-up in cohort studies. *American Journal of Epidemiology, 121,* 611-619.

Chesher, A., & Lancaster, T. (1983). The estimation of models of labour market behaviour. *Review of economic studies. 50,* 609-624.

Choy, S. P., Medrich, E. A., Henke, R. R., & Bobbitt, S. A. (1992). *Schools and Staffing in the United States: A Statistical Profile.* Washington, DC: National Center for Education Statistics.

Cook, N. R., & Ware, J. H. (1983). Design and analysis methods for longitudinal research, *Annual Review of Public Health, 4,* 1-23.

Darling-Hammond, L. (1984). *Beyond the commission reports: The coming crisis in teaching.* Santa Monica, CA: Rand Corporation.

de Stavola, B. L. (1986). Sampling designs for short panel data. *Econometrica, 54,* 415-424.

Duncan, G J., Juster, F. T., & Morgan, J. N. (1985). The role of panel studies in a world of scarce resources. In R. W. Pearson and R. F. Boruch (eds). *Survey research designs: Towards a better understanding of their costs and benefits*. New York, NY: Springer Verlag, 95-119.

Duncan, G. J., & Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review, 55,* 97-117.

Featherman, D. L. (1980). Retrospective longitudinal research: Methodological considerations. *Journal of Economics and Business, 32,* 152-169.

Fienberg, S. E., & Tanur, J. M. (1985). The design and analysis of longitudinal surveys: Controversies and issues of cost and continuity. In R. W. Pearson and R. F. Boruch (eds). *Survey research designs: Towards a better understanding of their costs and benefits*. New York, NY: Springer Verlag, 60-93.

Fienberg, S. E., & Tanur, J. M. (1987). Experimental and sampling structures: Parallels diverging and meeting. *International Statistical Review, 55* 75-96.

Friedman, W. J. (1993). Memory of time of past events. *Psychological Bulletin, 113,* 44-66.

Goldstein, H. (1979). *The design and analysis of longitudinal studies: Their role in the measurement of change*. New York: Academic Press.

Grissmer, D. W. & Kirby, S. N. (1987). *Teacher Attrition: The Uphill Climb to Staff the Nation's Schools*, Santa Monica, CA: Rand Corporation.

Grissmer, D. W., & Kirby, S. N. (1992). *Patterns of Attrition Among Indiana Teachers: 1965-1987,* Santa Monica, CA: Rand Corporation.

Haffner, A., & Owings, J. (1991). Careers in teaching: Following members of the high school class of 1972 in and out of teaching. Washington, DC: National Center for Education Statistics.

Haggstrom, G. W., Darling-Hammond, L., & Grissmer, D. W. (1988). *Assessing Teacher Supply and Demand*, Santa Monica, CA: Rand Corporation.

Heyns, B. (1988). Educational defectors: A first look at teacher attrition in the NLS-72, *Educational*

*Researcher, 17*, 24-32.

Hoem, J. M. (1985). Weighting, misclassification, and other issues in the analysis of survey samples of life histories. In J. J. Heckman and B. Singer (eds.) *Longitudinal analysis of labor market data.* New York, NY: Cambridge University Press.

Hogan, D. P. (1984). Cohort comparisons in the timing of life events. *Developmental review, 4* 289-310.

Holt, D. & Skinner, C. J. (1989). Components of change in repeated surveys. *International Statistical Review, 87*, 1-18.

Janson, C-G. (1981). Some problems of longitudinal research in the social sciences. In F. Schulsinger, S. A. Mednick, and J. Knop (eds.) *Longitudinal research: Methods and uses in Behavioral Science.* Boston, MA: Martinus Nijhoff Publishing.

Kennedy, M. M. (1992). The problem of improving teacher quality while balancing supply and demand. In E. E. Boe and D. M. Gilford (Eds.) *Teacher supply, demand, and quality: Policy issues, models, and data bases.* (pp. 65-108). Washington, DC: National Academy of Sciences Press.

Kish, L. (1986). *Statistical design for research.* New York, NY: Wiley.

Lancaster, T. (1990). *The Econometric Analysis of Transition Data.* Econometric Society Monographs, New York, NY: Cambridge University Press.

Magnusson, D., & Bergman, L. R. (1990). *Data quality in longitudinal research.* Cambridge, UK: Cambridge University Press.

Mason, W. M., & Fienberg, S. E. (1985). *Cohort Analysis in Social Research: Beyond the Identification Problem.* New York, NY: Springer-Verlag.

Means, B., Swan, G. E., Jobe, J. B., & Esposito, J. L. (1991). An alternative approach to obtaining personal history data. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman (eds.) *Measurement Errors in Surveys*. New York, NY: Wiley, 167-183.

Murnane, R. J., Singer, J. D., & Willett, J. B. (1988). The career paths of teachers: Implications for teacher supply and methodological lessons for research, *Educational Researcher*, 17, 22-30.

Murnane, R. J., Singer, J. D., & Willett, J. B. (1989). The Influences of Salaries and Opportunity Costs on Teachers' Career Choices: Evidence from North Carolina, *Harvard Educational Review, 59*, 325-346.

Murnane, R. J., Singer, J. D., Willett, J. B., Kemple, J. J., & Olsen, R. J. (1991). *Who Will Teach?: Policies that Matter*. Cambridge, MA: Harvard University Press.

Nesselroade, J., & Baltes, P. (1979). *Longitudinal research in the study of behavior and development.* New York: Academic Press.

Pearson, R. W. (1989). The advantages and disadvantages of longitudinal surveys. *Research in the sociology of education and socialization. 8*, 177-179.

Rogosa, D. R., Brand, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 90*, 726-748.

Schlesselman, J. J. (1973). Planning a longitudinal study: Frequency of measurement and study duration. *Journal of Chronic Diseases, 26*, 561-570.

Shulman, L. S. (1992). Directions for the future. In E. E. Boe and D. M. Gilford (Eds.) *Teacher supply, demand, and quality: Policy issues, models, and data bases.* (pp. 287-290). Washington, DC: National Academy of Sciences Press.

Silberstein, A R., & Scott, S. (1991). Expenditure diary surveys and their associated errors. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman (eds.) *Measurement Errors in Surveys*. New York, NY: Wiley, 303-326.

Singer, J. D., & Willett, J. B. (1991). Modeling the days of our lives: Using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. *Psychological Bulletin, 110*, 268-290.

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics, 18, 155-195.*

Theobold, N. (1990). An examination of the influence of personal, prefoessional, and school district

characteristics on public school teacher retention. *Economics of Education Review, 9*, 241-250.

Wagenaar, W. A. (1986). My memjory: A study of autobiographical memory over six years. *Cognitive Psychology, 18*, 225-252.

Wall, W. D., & Williams, H. L. (1970). *Longitudinal studies and the social sciences.* London: Heinemann.

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Rothkopf (Ed.). *Review of research in education* (1988-89( (pp. 345-422). Washington, DC: American Educational Research Association

Willett, J. B. & Singer, J. D. (1989). Two types of question about time: Methodological issues in the analysis of teacher career path data, *International Journal of Educational Research, 13*, 421-437.

Willett, J. B., & Singer, J. D. (1991). From whether to when: New methods for studying student dropout and teacher attrition, *Review of Educational Research, 61*, 407-450.

Willett, J. B., & Singer, J. D. (1993). Investigating onset, cessation, relapse, and recovery: Why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *Journal of Consulting and Clinical Psychology, 61*, 952-965.

Willett, J. B., & Singer, J. D. (in press). It's deja-vu all over again: Using multiple-spell discrete-time survival analysis, *Journal of Educational Statistics.*

# Listing of NCES Working Papers to Date

| Number | Title | Contact |
|---|---|---|
| 94-01 | Schools and Staffing Survey (SASS) Papers Presented at Meetings of the American Statistical Association | Dan Kasprzyk |
| 94-02 | Generalized Variance Estimate for Schools and Staffing Survey (SASS) | Dan Kasprzyk |
| 94-03 | 1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report | Dan Kasprzyk |
| 94-04 | The Accuracy of Teachers' Self-reports on their Postsecondary Education: Teacher Transcript Study, Schools and Staffing Survey | Dan Kasprzyk |
| 94-05 | Cost-of-Education Differentials Across the States | William Fowler |
| 94-06 | Six Papers on Teachers from the 1990-91 SASS and Other Related Surveys | Dan Kasprzyk |
| 94-07 | Data Comparability and Public Policy: New Interest in Public Library Data Papers Presented at Meetings of the American Statistical Association | Carrol Kindel |
| 95-01 | Schools and Staffing Survey: 1994 papers presented at the 1994 Meeting of the American Statistical Association | Dan Kasprzyk |
| 95-02 | QED Estimates of the 1990-91 Schools and Staffing Survey: Deriving and Comparing QED School Estimates with CCD Estimates | Dan Kasprzyk |
| 95-03 | Schools and Staffing Survey: 1990-91 SASS Cross-Questionnaire Analysis | Dan Kasprzyk |

## Listing of NCES Working Papers to Date (Continued)

| Number | Title | Contact |
|---|---|---|
| 95-04 | National Education Longitudinal Study of 1988: Second Follow-up Questionnaire Content Areas and Research Issues | Jeffrey Owings |
| 95-05 | National Education Longitudinal Study of 1988: Conducting Trend Analyses of NLS-72, HS&B, and NELS:88 Seniors | Jeffrey Owings |
| 95-06 | National Education Longitudinal Study of 1988: Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data | Jeffrey Owings |
| 95-07 | National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts | Jeffrey Owings |
| 95-08 | CCD Adjustments to the 1990-91 SASS: A Comparison of Estimates | Dan Kasprzyk |
| 95-09 | The Results of the 1993 Teacher List Validation Study (TLVS) | Dan Kasprzyk |
| 95-10 | The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation | Dan Kasprzyk |
| 95-11 | Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work | Sharon Bobbitt & John Ralph |
| 95-12 | Rural Education Data User's Guide | Samuel Peng |

## Listing of NCES Working Papers to Date (Continued)

| Number | Title | Contact |
|---|---|---|
| 95-13 | Assessing Students with Disabilities and Limited English Proficiency | James Houser |
| 95-14 | Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys | Samuel Peng |
| 95-15 | Classroom Instructional Processes: A Review of Existing Measurement Approaches and Their Applicability for the Teacher Follow-up Survey | Sharon Bobbitt |
| 95-16 | Intersurvey Consistency in NCES Private School Surveys | Steven Kaufman |
| 95-17 | Estimates of Expenditures for Private K-12 Schools | Steve Broughman |
| 95-18 | An Agenda for Research on Teachers and Schools: Revisiting NCES' Schools and Staffing Survey | Dan Kasprzyk |
| 96-01 | Methodological Issues in the Study of Teachers' Careers: Critical Features of a Truly Longitudinal Study | Dan Kasprzyk |